On the relation between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment

Nobuaki Hoshino
Graduate School of Economics, University of Tokyo
and
Akimichi Takemura
Faculty of Economics, University of Tokyo

March, 1998

Abstract

Fisher's logarithmic series model (Fisher et al. (1943)) is a classical model in statistical ecology. In this paper we show that this model is a key model linking three models discussed in Takemura (1997), i.e., Poisson-gamma model (Bethlehem et al. (1990)), Dirichlet-multinomial model (Takemura (1997)), and Ewens model (Ewens (1990)). This connection opens up the possibility of applying existing techniques of statistical ecology to the problem of microdata disclosure risk assessment.

1 Introduction

Fisher's logarithmic series model was the starting point of the whole area of statistical ecology. The logarithmic series model and other related models for populations consisting of large number of species are called "abundance models" or "stochastic abundance models" (Engen (1978)). In disclosure risk assessment of categorical microdata set, we often find that the total number of cells of the population is very large, because it is the product of the number of categories of many categorical or categorized attributes. Therefore the models developed in statistical ecology are likely to be useful in disclosure risk assessment of microdata sets.

In releasing a microdata set, the statistical agency can control the coarseness of the categorization ("global recoding") of each attribute. Detailed categorization of each variable decreases the cell size (the number of individuals falling in the cell) and the identification risk of individuals of a cell is thought to be inversely related to the cell size. Thus microdata sets with many small cells are regarded as risky. In particular, the population uniques may be identified when their attributes are disclosed. The number or the proportion of the population uniques is one of the key quantities in disclosure risk assessment of a microdata set.

Bethlehem et al. (1990) proposed the Poisson-gamma model for estimating the number of population uniques. For a survey on the Poisson-gamma model see Skinner (1992). In the Poisson-gamma model the population size N is a random variable having a negative binomial distribution. Takemura (1997) derived Dirichlet-multinomial model as conditional model where the population size N of the Poisson-gamma model is fixed. Equivalently the Poisson-gamma model is obtained by taking mixture of the population size N of the Dirichlet-multinomial model according to negative binomial distribution. In this paper we will prove that the same relation holds between the Ewens model and the logarithmic series model, i.e., the Ewens model is a conditional model of the logarithmic series model where the population size N is fixed and the logarithmic series model is obtained by taking mixture of the population size N of the Ewens model according to negative binomial distribution.

Takemura (1997) also showed that the Ewens model is obtained from the Dirichlet-multinomial model by a limiting argument similar to the law of small numbers. We will prove that the logarithmic series model can be obtained from the Poisson-gamma model by exactly the same limiting argument. The above two parallel relations can be conveniently summarized by saying that the conditioning and the limiting process commute in Figure 1.

Furthermore we show that simple random sampling without replacement from the logarithmic series model leads to the same sampling distribution as simple random sampling without replacement from the Ewens model. Therefore under simple random sampling without replacement, inference on the logarithmic series model is exactly the same as the inference on the Ewens model.

In Section 2 we set up appropriate notation and prove our main results. In Section 3 we discuss sampling from the logarithmic series model.

2 Main results

Consider a discrete population of size N. Let K denote the total number of the cells and let $F_j,\ j = 1, \ldots, K$, denote the size of the j-th cell. In superpopulation model we consider $F_j,\ j = 1, \ldots, K$, as random variables. Let S_i denote the number of cells of size i. In terms of the indicator function

$$I(F_j = i) = \begin{cases} 1, & F_j = i, \\ 0, & F_j \neq i, \end{cases}$$

S_i can be expressed as

$$S_i = \sum_{j=1}^{K} I(F_j = i).$$

$S_i,\ i = 0, 1, \ldots,$ are called size indices (Sibuya (1993)) or frequencies of frequencies (Good (1965)). Obviously

$$\sum_{i=0}^{\infty} S_i = K, \qquad \sum_{i=1}^{\infty} i \cdot S_i = N.$$

In microdata disclosure risk assessment, the number of population uniques S_1 or the proportion of population uniques $\pi = S_1/N$ are of particular importance.

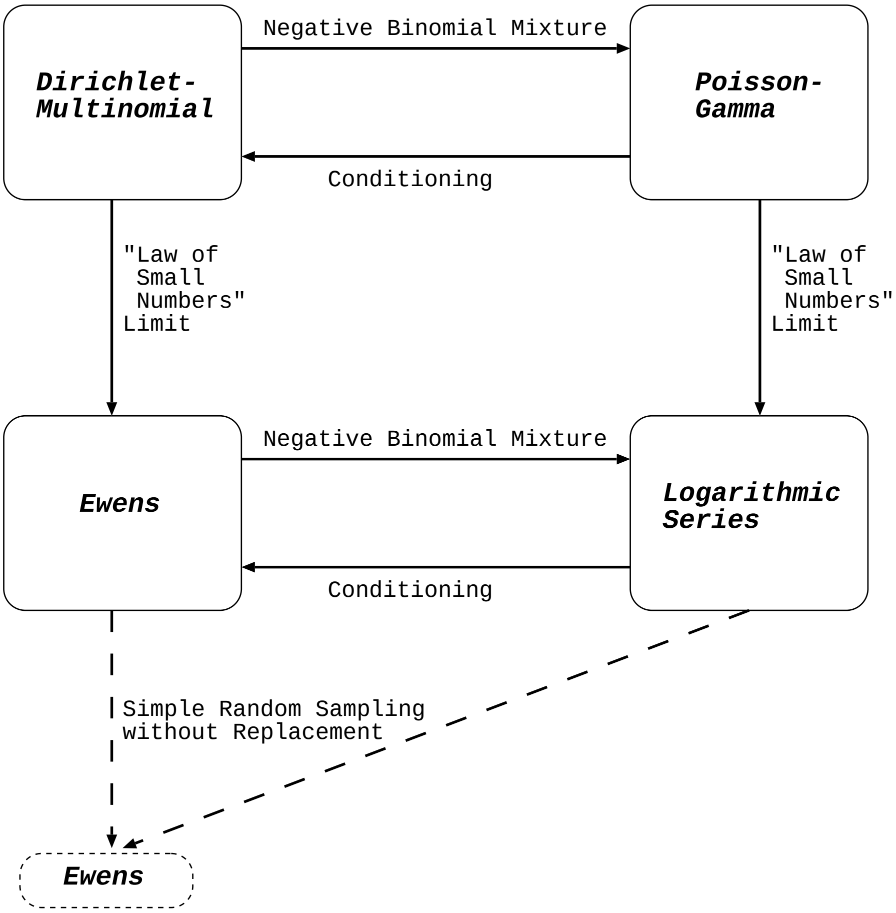
Negative Binomial Mixture
Dirichlet-Multinomial
Poisson-Gamma
Conditioning
"Law of Small Numbers" Limit
"Law of Small Numbers" Limit
Negative Binomial Mixture
Ewens
Logarithmic Series
Conditioning
Simple Random Sampling without Replacement
Ewens

Figure 1: Diagram of four models

Fisher's logarithmic series model is defined in terms of the joint distribution of size indices S_i, $i \geq 1$. Empty cells are not observed and S_0 is not defined in this model. Let $S_1, S_2, \ldots,$ be independent Poisson random variables with mean

$$\mathrm{E}(S_i) = \lambda_i = N_0 \frac{p \cdot q^{i-1}}{i}, \quad i = 1, 2, \ldots,$$

where $N_0 > 0,\ 0 < p < 1$ and $q = 1 - p$. The joint probability function of the size indices $(S_1, S_2, \ldots)$ is given as

$$\mathrm{P}(S_1, S_2, \ldots) = \prod_{i=1}^{\infty} \frac{\lambda_i^{S_i} \exp(-\lambda_i)}{S_i!}, \tag{1}$$

where only finite number of S_i's are nonzero. Note that

$$\mathrm{E}(N) = \sum_{i=1}^{\infty} i\mathrm{E}(S_i) = \sum_{i=1}^{\infty} i\lambda_i = N_0 p \sum_{i=1}^{\infty} i \cdot \frac{q^{i-1}}{i} = N_0$$

is the expected population size.

We first derive this model from the Poisson-gamma model of Bethlehem et al. (1990) by a limiting argument similar to the law of small numbers. In the Poisson-gamma model F_j is a Poisson random variable with mean $N_0\mu$ and μ has gamma distribution with parameters α, β. The unconditional distribution of F_j is negative binomial distribution with the following probability function.

$$\begin{aligned} \mathrm{P}(F_j = i) &= \int_0^{\infty} \frac{(N_0\mu)^i \exp(-N_0\mu)}{i!} \frac{\mu^{\alpha-1}\exp(-\mu/\beta)}{\Gamma(\alpha)\beta^{\alpha}} d\mu \\ &= \frac{\Gamma(i+\alpha)}{\Gamma(\alpha) i!} p^{\alpha} q^{i}, \qquad (2) \end{aligned}$$

where $q = N_0\beta/(1+N_0\beta)$, $p = 1-q$. In the Poisson-gamma model α, β are assumed to satisfy the restriction $\alpha\beta = 1/K$. Furthermore F_j, $j = 1, \dots, K$, are assumed to be independently and identically distributed. In summary the Poisson-gamma model is defined by the joint probability function of F_j's as

$$\mathrm{P}(F_1, \dots, F_K) = \prod_{j=1}^{K} \frac{\Gamma(F_j + \alpha)}{\Gamma(\alpha) F_j!} p^{\alpha} q^{F_j}, \qquad q = \frac{N_0\beta}{N_0\beta + 1}\ ,\ p = 1 - q,\ \alpha\beta = \frac{1}{K}. \qquad (3)$$

In this model $N = F_1 + \cdots + F_K$ is a random variable having negative binomial distribution with α replaced by $A = K\alpha$ in (2), because negative binomial distribution is closed under convolution. $\mathrm{E}(N) = K\mathrm{E}(F_j) = KN_0\alpha\beta = N_0$ is the expected population size.

Takemura (1997) showed that the conditional Poisson-gamma model given N equals the Dirichlet-multinomial model defined by

$$\mathrm{P}(F_1, \dots, F_K \mid N) = \frac{N!\Gamma(K\alpha)}{\Gamma(K\alpha + N)} \prod_{j=1}^{K} \frac{\Gamma(\alpha + F_j)}{\Gamma(\alpha) F_j!}.$$

Let $K\alpha = A$ be fixed and consider a limiting process similar to the law of small numbers:

$$K \to \infty, \quad \alpha \to 0, \qquad (A = K\alpha : \text{fixed}). \qquad (4)$$

Here N and $\beta = 1/(K\alpha)$ remain to be fixed. Takemura (1997) showed that the marginal distribution of $(S_1, S_2, \dots)$ of the Dirichlet-multinomial model converges to the Ewens model with parameter A:

$$\mathrm{P}(S_1, \dots, S_N) = \frac{A^U}{A^{[N]}} \frac{N!}{\prod_{i=1}^{N} i^{S_i} S_i!}, \qquad (5)$$

where $A^{[N]} = A(A+1)(A+2)\cdots(A+N-1)$, $U = \sum_{i=1}^{N} S_i$.

We now apply the same limiting process (4) to the Poisson-gamma model. The following result was already implicit in Fisher et al. (1943). See Anscombe (1950), Section 3.2 of Engen (1978) or Johnson et al. (1993) for more discussion.

Proposition 1 *Apply the limiting process* (4) *to the Poisson-gamma model in* (3)*. Then the marginal distribution of* $(S_1, S_2, \ldots)$ *converges to Fisher's logarithmic series model in* (1)*.*

For convenience we give a proof of this proposition in Appendix.

Now we consider fixing the value of N in the logarithmic series model. As already mentioned N has negative binomial distribution

$$\begin{aligned} \mathrm{P}(N) &= \frac{\Gamma(K\alpha + N)}{\Gamma(K\alpha)N!} p^{K\alpha} q^N \\ &= \frac{\Gamma(A+N)}{\Gamma(A)N!} p^A q^N. \end{aligned} \tag{6}$$

Note that this distribution does not change by the limiting process in (4). Therefore $N = \sum_i iS_i$ has the same negative binomial distribution under the logarithmic series model. More precisely

$$\sum_{S \in \mathcal{S}_N} \prod_{i=1}^{\infty} \frac{\lambda_i^{S_i} \exp(-\lambda_i)}{S_i!} = \frac{\Gamma(A+N)}{\Gamma(A)N!} p^A q^N,$$

where $S = (S_1, S_2, \ldots)$ and

$$\mathcal{S}_N = \{S \mid \sum_i iS_i = N\}.$$

Then

$$\begin{aligned} \mathrm{P}(S_1, S_2, \ldots \mid N) &= \{\prod_i \frac{\lambda_i^{S_i} \exp(-\lambda_i)}{S_i!}\} \,/\, \{\frac{\Gamma(A+N)}{\Gamma(A)N!} p^A q^N\} \\ &= \prod_i \frac{(\frac{A}{i} q^i)^{S_i} \exp(-\lambda_i)}{S_i!} \cdot \frac{N!}{A^{[N]}} p^{-A} q^{-N} \\ &= \frac{N!}{A^{[N]}} p^{-A} q^{-N} \cdot A^U \exp(-\sum_i \lambda_i) q^N \prod_i \frac{1}{i^{S_i} S_i!} \\ &= \frac{A^U N!}{A^{[N]}} p^{-A} \cdot \exp(-\sum_i \lambda_i) \prod_i \frac{1}{i^{S_i} S_i!}, \end{aligned}$$

where $U = \sum_{i=1}^N S_i$. By

$$\sum_{i=1}^{\infty} \lambda_i = A \sum_{i=1}^{\infty} \frac{q^i}{i} = -A \log(1-q),$$

the conditional distribution is

$$\begin{aligned} \mathrm{P}(S_1, S_2, \ldots \mid N) &= \frac{A^U N!}{A^{[N]}} p^{-A} \cdot \exp(A \log(1-q)) \prod_{i=1}^{\infty} \frac{1}{i^{S_i} S_i!} \\ &= \frac{A^U N!}{A^{[N]}} p^{-A} \cdot p^A \prod_{i=1}^{\infty} \frac{1}{i^{S_i} S_i!} \\ &= \frac{A^U N!}{A^{[N]}} \prod_{i=1}^{\infty} \frac{1}{i^{S_i} S_i!}, \end{aligned} \tag{7}$$

and this equals (5). We have proved that the conditional logarithmic series model given N coincides with the Ewens model with parameter A. Conversely by the product of (7) and (6) it can be easily proved that if N is distributed according to negative binomial distribution then the mixture of the Ewens model becomes the logarithmic series model.

Theorem 1 *The conditional logarithmic series model given N is the Ewens model with parameter $A = K\alpha = 1/\beta$. Conversely if N has negative binomial distribution of* (6) *in the Ewens model, then the mixture becomes the logarithmic series model.*

3 Sampling from logarithmic series model

Here we consider sampling from the logarithmic series model and estimation of the number S_1 or the proportion π of population uniques from the sample. We discuss two sampling schemes: 1) simple random sampling without replacement, 2) the Bernoulli sampling. Sample size indices are denoted by $s_1, s_2, \ldots, s_n$, where n is the sample size.

3.1 Simple random sampling without replacement

From Theorem 1 and the fact that the Ewens model is closed under simple random sampling without replacement (see Section 4 of Takemura (1997)), we have the following proposition.

Proposition 2 *Let $(s_1, s_2, \ldots, s_n)$ be the sample size indices obtained by simple random sampling without replacement from the logarithmic series model. Then $(s_1, s_2, \ldots, s_n)$ has the Ewens distribution in* (5) *with N replaced by n.*

Proof is the same as in Corollary 1 of Takemura (1997) and omitted.

Maximum likelihood estimation on the Ewens model is discussed by Sibuya (1991). We rewrite the result in our notation. By (7) log likelihood of the sample is

$$L = u \log A - \log(A + n - 1) - \log(A + n - 2) - \cdots - \log A + \text{const},$$

where $u = \sum_{i=1}^{n} s_i$. Thus the maximum likelihood estimator of A is the solution of

$$\frac{u}{A} - \sum_{j=1}^{n} \frac{1}{A - 1 + j} = 0.$$

This is easily solved by the Newton-Raphson method with starting value $A = s_1$.

By Sibuya (1993)

$$\mathrm{E}(S_i) = \frac{A}{i} \prod_{j=1}^{i} \frac{N - j + 1}{A + N - j}.$$

In particular

$$\mathrm{E}(s_1) = A \frac{n}{A + n - 1} \quad \text{and} \quad \mathrm{E}(S_1) = A \frac{N}{A + N - 1}.$$

Therefore a simple moment estimator of A is

$$\hat{A} = \frac{(1-n)s_1}{s_1 - n}$$

and the proportion of population uniques can be estimated by

$$\hat{\pi} = \frac{s_1(1-n)}{s_1(N-n) - n(N-1)}.$$

3.2 Bernoulli sampling

In the Bernoulli sampling, a coin with success probability r is tossed for each individual and the individual is drawn if the coin results in heads. See Section 2.2 of Särndal et al. (1992) or Appendix A of Takemura (1997). The Bernoulli sampling is only a convenient approximation to simple random sampling without replacement in the context of sampling of official statistics. However in ecological sampling, it is more natural than simple random sampling without replacement. For example, r may represent the probability of each animal caught in a trap. Under the Bernoulli sampling the sampling distribution of the logarithmic series model is easily derived.

Suppose that the sample is obtained by the Bernoulli sampling with $r = n_0/N_0$. For convenience we replace n_0/N_0 by n/N. Then

$$\mathrm{P}(s_1, s_2, \ldots) = \prod_{i=1}^{\infty} \frac{\tilde{\lambda}_i^{s_i} \exp(-\tilde{\lambda}_i)}{s_i!},$$

where $\tilde{\lambda}_i = npq^{i-1}/i$. The log likelihood function turns out to be

$$L = \frac{r}{\beta} \log p - u \log \beta + n \log q + \text{const},$$

where $u = \sum_{i=1}^{n} s_i$. Differentiating L by β, the maximum likelihood estimate $\hat{\beta}$ is the solution of

$$n \log(N\beta + 1) - uN\beta = 0.$$

This can be solved by the Newton-Raphson method. Then the proportion of population uniques is estimated as

$$\hat{\pi} = \hat{p} = \frac{1}{1 + N\hat{\beta}}.$$

If the sample size is sufficiently large then $\hat{\beta}$ will give a reasonable estimate for simple random sampling without replacement.

Appendix A : Proof of Proposition 1

Consider the joint probability generating function of $S_1, S_2, \ldots,$ under the Poisson-gamma model. If $K = 1$, then

$$\begin{aligned}
G_1(z_1, z_2, \ldots) &= \mathrm{E}[\prod_{i=1}^{\infty} z_i^{S_i}] \\
&= z_1 \mathrm{P}(S_1 = 1, S_2 = 0, S_3 = 0, \ldots) \\
&\quad + z_2 \mathrm{P}(S_1 = 0, S_2 = 1, S_3 = 0, \ldots) \\
&\quad + \cdots \\
&\quad + 1 \times \mathrm{P}(S_0 = 1, S_1 = 0, S_2 = 0, \ldots) \\
&= \sum_{i=1}^{\infty} z_i \mathrm{P}(\ldots, S_i = 1, \ldots) + 1 \times (1 - \sum_{i=1}^{\infty} \mathrm{P}(\ldots, S_i = 1, \ldots)) \\
&= \sum_{i=1}^{\infty} (z_i - 1) \mathrm{P}(\ldots, S_i = 1, \ldots) + 1 \\
&= \sum_{i=1}^{\infty} (z_i - 1) \mathrm{P}(F_1 = i) + 1 \\
&= \sum_{i=1}^{\infty} (z_i - 1) \frac{\Gamma(i+\alpha)}{\Gamma(\alpha) i!} p^{\alpha} q^i + 1.
\end{aligned}$$

By the independence of F_j's, the joint probability generating function for general K is expressed as

$$\begin{aligned}
G(z_1, z_2, \ldots) &= G_1(z_1, z_2, \ldots)^K \\
&= \Big[\sum_{i=1}^{\infty} (z_i - 1) \frac{\Gamma(i+\alpha)}{\Gamma(\alpha) i!} p^{\alpha} q^i + 1\Big]^K.
\end{aligned} \tag{8}$$

Now consider the limiting process (4). Then the limit of (8) is

$$\begin{aligned}
&[\sum_{i=1}^{\infty} (z_i - 1) \frac{(i+\alpha-1)(i+\alpha-2)\cdots(\alpha+1)\alpha}{i!} p^{\alpha} q^i + 1]^K \\
&= [1 + \frac{1}{K} \sum_{i=1}^{\infty} (z_i - 1) \frac{(i+\alpha-1)(i+\alpha-2)\cdots(\alpha+1)}{i!} A p^{\alpha} q^i]^K \\
&\to \exp(\sum_{i=1}^{\infty} (z_i - 1) \frac{(i-1)(i-2)\cdots 1}{i!} A q^i) \\
&= \exp(\sum_{i=1}^{\infty} (z_i - 1) \frac{A}{i} q^i).
\end{aligned}$$

Using $A = K\alpha = 1/\beta$ and $Aq^i = Aq \cdot q^{i-1} = N_0 p q^{i-1}$, the limiting joint probability generating function can be written as

$$\lim_{K \to \infty} G(z_1, z_2, \ldots) = \exp(\sum_{i=1}^{\infty} \lambda_i (z_i - 1)),$$

where

$$\lambda_i = \frac{N_0 \cdot p \cdot q^{i-1}}{i}.$$

This agrees with the joint probability generating function of independent Poisson variables $S_i, i = 1, 2, \ldots,$ with mean $\mathrm{E}(S_i) = \lambda_i$.

References

[1] Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**, 358–382.

[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] Engen, S. (1978). *Stochastic Abundance Models.* Chapman and Hall, London.

[4] 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.

[5] Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.

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

[7] Johnson, N.L., Kotz, S. and Kemp, A.W. (1993). *Univariate Discrete Distributions*, 2nd ed., Chap. 7, Wiley, New York.

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

[9] Sibuya, M. (1991). A cluster-number distribution and its application to the analysis of homonyms. *Japanese Journal of Applied Statististics*, **20**, 139–153 (in Japanese).

[10] Sibuya, M. (1993). A random clustering process. *Annals of Institute of Statistical Mathematics*, **45**, 459–465.

[11] Skinner, C.J. (1992). On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, **46**,21–32.

[12] Takemura, A. (1997). Some superpopulation models for estimating the number of population uniques. Discussion Paper 97-F-29, Faculty of Economics, University of Tokyo.