Journal of Official Statistics, Vol. 17, No. 4, 2001, pp. 499–520

Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment

Nobuaki Hoshino[1]

Ewens's sampling formula (Ewens 1972), which is mainly studied in statistical ecology, has been used to assess the microdata disclosure risk. Pitman (1995) considered an extension of the Ewens sampling formula, and in the present article we evaluate the usefulness of the Pitman sampling formula in the disclosure field. First we clarify some theoretical implications of the Pitman model as a tool for assessing the risk. We then compare various models based on the Akaike Information Criterion (AIC) by applying them to real data sets from the Japanese labor force survey. Our comparison strongly supports the Pitman model. These results suggest that the Pitman sampling formula is very promising for the microdata disclosure problem as well as for statistical ecology.

Key words: Privacy; uniqueness; species abundance; superpopulation; random clustering.

1. Introduction

In releasing a microdata set, the statistical agency must exclude individual records that are identifiable to the individual. A record is composed of fields that correspond to categorized attributes of an individual. Attackers might identify an individual using information on records. In practical terms, we may consider individuals that are unique in the population with respect to the categorization in the sample data to be identifiable. The number of population uniques is thus an important control object in the context of microdata disclosure, and it is important to estimate the number of population uniques from sample data at hand. After estimating the number of population uniques, we may regard the number of the uniques times the sampling ratio as the estimate of the number of population uniques contained in the sample. It is unsafe to disseminate a data set that contains many population uniques.

To estimate population uniques, Bethlehem et al. (1990) introduced the Poisson-gamma model, which is the first application of the superpopulation model in the field of the microdata disclosure problem. Under the superpopulation model based approach, we assume that the population is generated by an appropriate (prior) distribution. By means of the assumption on the prior distribution, the risk inference is reduced to the problem of parameter estimation. We should be pragmatic since it is impossible to know the

[1] Faculty of Economics, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp
Acknowledgments: The author would like to thank Professor Akimichi Takemura for many useful comments and Professor Masaaki Sibuya for stimulating discussions.

true mechanism of generating population. Here we adopt the empirical Bayes method; what is required is a prior distribution flexible enough to describe various populations.

We briefly survey various superpopulation models used in the literature. Several authors apply the Poisson-gamma model to actual data sets, but reported fits are bad. See Skinner (1992) or Skinner and Holmes (1993). Skinner and Holmes (1993) applied the Poisson-lognormal model and the logarithmic series distribution to U.S. and Italian data sets. These models have mainly been studied in ecology, where frequencies of species are estimated from sample frequency structure. The stochastic abundance model (Engen 1978) is used for modeling the populations consisting of large numbers of species in statistical ecology. Hoshino and Takemura (1998) clarified relations between various superpopulation models and revealed that the superpopulation model based approach in respect of the disclosure problem has a connection with the stochastic abundance models. The Poisson-lognormal model is studied, for example, in Bulmer (1974) and in Aitchison and Ho (1989). Fisher's classical logarithmic series model (Fisher et al. 1943) leads to many versions of superpopulation models; see Section 3.2 of Engen (1978) and Johnson et al. (1993). Hoshino and Takemura (1998) noted, on the basis of an interpretation of Anscombe (1950), that a limiting Poisson-gamma model becomes a logarithmic series model different from that of Skinner and Holmes (1993). Takemura (1999) considered a sampling distribution from the Poisson-gamma model and derived the Dirichlet-multinomial model. Takemura (1999) also identified that the Ewens sampling formula originally developed in genetics is a limiting form of the Dirichlet-multinomial model. See Ewens (1990), Sibuya (1993), and Johnson et al. (1997) for the Ewens distribution. In Hoshino and Takemura (1998), we showed that the Ewens model is derived from the logarithmic series model by the same conditioning argument as the Dirichlet-multinomial model is derived from the Poisson-gamma model. Watterson (1973) referred to the Ewens distribution as a version of the logarithmic series distribution. Samuels (1998) applied the Ewens distribution to real microdata sets. However, it is stated that the fitted parameter values were too small to provide proper risk inference.

Pitman (1995) considered the random partition of the positive integers, and obtained a new generalization of the Ewens distribution. See Pitman (1996), Pitman and Yor (1997), and Yamato et al. (1999) for the context. The obtained distribution is the Pitman sampling formula. Samuels (1998) negatively mentioned the Pitman model, because the objective was to set larger parameter values for the Ewens model. We will later show that the Pitman model is not suitable for this purpose. However, this model is flexible in nature; it contains the Ewens model and the Dirichlet-multinomial model as special cases. Thus the fit of the Pitman model is at least as good as those of these models, although the degree of freedom decreases.

If the Pitman model greatly improves prediction of the disclosure risk, then the superpopulation model based approach becomes much more relevant not only for the disclosure problem but also for the stochastic abundance model fitting. It is important to apply the above superpopulation models to actual data sets and compare each model on the same appropriate criterion.

The present article treats the method of applying the Pitman model to the disclosure problem and provides a comparison of the Pitman model with the other models. In Section 2 we derive some relevant moments of the Pitman model. Estimation with the

Pitman model is discussed in Section 3. Section 4 demonstrates the applicability of the Pitman model. After discussing model selection policy, we compare the Pitman model with other superpopulation models by applying them to Japanese labor force survey data sets. Section 5 offers motivation for the Pitman model and some concluding remarks. In the rest of this section we provide notation and define the existing superpopulation models compared in Section 4.

1.1. Notation and summary of existing superpopulation models

Consider a discrete population of size N. Let K denote the total number of cells and let F_j, $j = 1, \ldots, K$, denote the size of the j-th cell. Under the superpopulation model approach we consider F_j, $j = 1, \ldots, K$, as random variables; the population size $N = \sum_{j=1}^{K} F_j$ may or may not be a random variable. Let S_i denote the number of cells of size i. In terms of the indicator function

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

the number of cells of size i is expressed as

$$S_i = \sum_{j=1}^{K} I(F_j = i),\ i = 0, 1, \ldots,$$

which are called size indices (Sibuya (1993)) or frequencies of frequencies (Good 1965). These ideas correspond to equivalence class (Greenberg and Zayatz (1992)) in the context of the disclosure problem. In the risk assessment, the number of population uniques S_1 is of particular importance.

Obviously

$$\sum_{i=0}^{\infty} S_i = K \quad \sum_{i=1}^{\infty} i \cdot S_i = N$$

Here K is the total number of cells including the number of the empty cells S_0. In the following we denote the number of nonempty cells by

$$U = K - S_0 = \sum_{i=1}^{\infty} S_i$$

One important difference between the disclosure problem and statistical ecology is the handling of U and K. In statistical ecology we usually only consider the marginal distribution of $(S_1, \ldots)$, and do not include K in the models. The reason is that species of frequency zero in a population have little meaning and there is no obvious means to specify S_0 in statistical ecology. However, as far as the microdata problem is concerned, we can set K as the product of the number of categories in variables assessed. Generally K becomes huge. The limiting process of $K \to \infty$ is thus reasonable.

In the following we summarize existing superpopulation models. We classify these models by paying attention to the following two points: (a) whether the population size N is a random variable or not, and (b) whether S_0 is defined or not. Models in which S_0 is not defined are described without explicit dependence on K.

Poisson-gamma model: The population size N is a random variable having the negative binomial distribution, and S_0 is defined. Let X_{PG} be the Poisson random variable with mean $N_0\mu$, and μ has the gamma distribution with its density: $f(\mu) = \mu^{\gamma-1}\exp(-\mu/\beta)/(\Gamma(\gamma)\beta^{\gamma})$. The parameters γ and β are assumed to satisfy the restriction $\gamma\beta = 1/K$. The unconditional distribution of X_{PG} becomes the negative binomial distribution. Under the Poisson-gamma model, F_j, $j = 1, \ldots, K$, are assumed to be independently and identically distributed as X_{PG}. In summary the Poisson-gamma model is defined by the joint probability function of F_j's as

$$P(F_1, \ldots, F_K) = \prod_{j=1}^{K} \frac{\Gamma(F_j + \gamma)}{\Gamma(\gamma)F_j!}\, p^{\gamma} q^{F_j} \qquad q = \frac{N_0\beta}{N_0\beta + 1}, \quad p = 1 - q, \quad \gamma\beta = \frac{1}{K} \tag{1}$$

The expected population size is $\mathrm{E}(N) = K\mathrm{E}(F_j) = KN_0\gamma\beta = N_0$

Poisson-lognormal model: The population size N is a random variable, and S_0 is defined. As in the Poisson-gamma model, X_{PL} is the Poisson random variable with mean λ. In addition, $\log\lambda$ is normally distributed with mean M and variance V. We assume that F_j, $j = 1, \ldots, K$, are independently and identically distributed as X_{PL}. The Poisson-lognormal model is defined by

$$P(F_1, \ldots, F_K) = \prod_{j=1}^{K} \frac{1}{F_j!\sqrt{2\pi V}} \int_0^{\infty} \lambda^{F_j - 1} \exp(-\lambda - (\log\lambda - M)^2/2V)\, d\lambda \tag{2}$$

The expected population size becomes $K\exp(M + V/2)$. In the present article we restrict the model such that $K\exp(M + V/2) = N_0$. Thus $M = \log N_0 - \log K - V/2$, and V is the unique parameter.

Dirichlet-multinomial model: The population size N is fixed, and S_0 is defined. The Dirichlet-multinomial model is the conditional Poisson-gamma model given N and defined by

$$\mathrm{P}(S_0, \ldots, S_N) = \frac{N!K!\Gamma(K\gamma)}{\Gamma(K\gamma + N)} \prod_{i=0}^{N} \left(\frac{\Gamma(\gamma + i)}{\Gamma(\gamma)i!}\right)^{S_i} \frac{1}{S_i!} \tag{3}$$

Logarithmic series model: The population size N is a random variable, and S_0 is not defined. Fisher's logarithmic series model is defined in terms of the joint distribution of size indices S_i, $i \geq 1$. Let

$$\lambda_i = N_0 \frac{p \cdot q^{i-1}}{i}, \qquad i = 1, 2, \ldots$$

where $N_0 > 0$, $0 < p < 1$ and $q = 1 - p$. Here S_i is an independent Poisson random variable with mean λ_i. The joint probability function of the size indices $(S_1, S_2, \ldots)$ becomes

$$\mathrm{P}(S_1, S_2, \ldots) = \prod_{i=1}^{\infty} \frac{\lambda_i^{S_i} \exp(-\lambda_i)}{S_i!} \tag{4}$$

Here only a finite number of S_i's are nonzero. This model is the limiting form of the Poisson-gamma model as $K \to \infty$ with $K\gamma$ fixed.

Ewens model: The population size N is fixed, and S_0 is not defined. Applying the limiting process $K \to \infty$ to the Dirichlet-multinomial model with $K\gamma = 1/\beta = \theta$ fixed, we obtain the Ewens model with parameter θ:

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

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

2. Some Theoretical Results on the Pitman Sampling Formula

In this section we introduce the Pitman model and derive some moments of the model needed for the estimation.

Sibuya (1993) describes the urn scheme construction of the Ewens sampling formula. It is instructive to consider a similar urn model construction of the Pitman model. Let us consider the following random clustering process: suppose that n balls (individuals) are distributed over u urns (cells) such that no empty urn exists; the number of balls in the j-th urn is denoted by f_j; we then put the $n+1$-st ball into an urn; with the probability of

$$\frac{\theta + u\alpha}{\theta + n} \tag{6}$$

we put the ball into the $u+1$-st urn that was empty; otherwise the number of nonempty urns remains unchanged, and the ball is put into one of the nonempty u urns; the probability of choosing the j-th nonempty urn is

$$\frac{f_j - \alpha}{\theta + n} \tag{7}$$

where $j = 1, \ldots, u$. Starting the process with $n = u = 0$, we can inductively derive the Pitman sampling formula. See Proposition 9 of Pitman (1995), which gives a more formal description of the process.

For each pair of real parameters α and θ, such that either $0 \le \alpha < 1$ and $\theta > -\alpha$, or $\alpha < 0$ and $\theta = -m\alpha$ for some natural number m, the Pitman model is defined by

$$\mathrm{P}(S_1, \ldots, S_N) = N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=1}^{N} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{S_j} \frac{1}{S_j!} \tag{8}$$

where $\theta^{[U:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(U-1)\alpha)$, $\theta^{[N]} = \theta(\theta+1)\cdots(\theta+N-1)$. If α equals zero, (8) amounts to the Ewens model (5). Assuming that $\alpha < 0$, let $\theta = -K\alpha > 0$, $\gamma = -\alpha > 0$. Then (8) amounts to the Dirichlet-multinomial model (3).

Write

$$S_{N,U} = \{S = (S_1, \ldots, S_N) \Big| \sum_{i \ge 1} iS_i = N, \qquad \sum_{i \ge 1} S_i = U\}$$

In the following we consider $U_N = \sum_{i=1}^{N} S_i$ as a random variable given N. The equations (6) and (7) imply that

$$\begin{aligned} \mathrm{P}(U_{N+1} = U) = &\mathrm{P}(U_{N+1} = U | U_N = U)\mathrm{P}(U_N = U) \\ &+ \mathrm{P}(U_{N+1} = U | U_N = U-1)\mathrm{P}(U_N = U-1) \end{aligned} \tag{9}$$

and

$$\mathrm{P}(U_{N+1} = U+1 | U_N = U) = \frac{\theta + U\alpha}{\theta + N}, \quad \mathrm{P}(U_{N+1} = U | U_N = U) = \frac{N - U\alpha}{\theta + N} \tag{10}$$

Yamato et al. (1999) give the explicit form of $\mathrm{P}(U_N = U)$, although it is complicated.

Theorem 1 *Suppose that size indices are distributed according to (8). Then*

$$\mathrm{E}(S_i) = \frac{(1-\alpha)^{[i-1]}}{i!}(\theta + \alpha \cdot \mathrm{E}(U_{N-i})) \prod_{j=1}^{i} \frac{N-j+1}{\theta + N - j} \tag{11}$$

where

$$\mathrm{E}(U_{N-i}) = \frac{\theta}{\theta + N - i - 1} + \sum_{l=0}^{N-i-2} \frac{\theta}{\theta + l} \prod_{j=l+1}^{N-i-1} \left(1 + \frac{\alpha}{\theta + j}\right) \tag{12}$$

Proof First we show (11).

$$\begin{aligned} E(S_i) &= \sum_{U=1}^{N} \sum_{S \in S_{N,U}} N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=1}^{N} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{S_j} \frac{1}{S_j!} S_i \\ &= \sum_{U=1}^{N} \sum_{S \in S_{N,U}} N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=1(j \neq i)}^{N} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{S_j} \frac{1}{S_j!} \left(\frac{(1-\alpha)^{[i-1]}}{i!} \right)^{S_i} \frac{1}{(S_i - 1)!} \\ &= \sum_{U=1}^{N} \sum_{S \in S_{N-i,U-1}} N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=1}^{N} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{S_i} \frac{1}{S_j!} \frac{(1-\alpha)^{[i-1]}}{i!} \\ &= \sum_{U=1}^{N} \frac{(1-\alpha)^{[i-1]}}{i!} \prod_{l=1}^{i} \frac{N-l+1}{\theta + N - l} (\theta + (U-1)\alpha) \mathrm{P}(U_{N-i} = U-1) \end{aligned} \tag{13}$$

Since $\sum_{U=1}^{N}(U-1)\mathrm{P}(U_{N-i} = U-1) = \mathrm{E}(U_{N-i})$, (11) is proved.

To prove (12) we utilize a recurrence relation:

$$\mathrm{E}(U_{N+1)} = \mathrm{E}(U_N)\left(1 + \frac{\alpha}{\theta + N}\right) + \frac{\theta}{\theta + N} \tag{14}$$

where $\mathrm{E}(U_0) = 0$. Assuming that (14) is true, we can easily prove (12) by induction.

Using (9) and (10),

$$\begin{aligned} \mathrm{E}(U_{N+1} | U_N = U) &= U\mathrm{P}(U_{N+1} = U | U_N = U) + (U+1)\mathrm{P}(U_{N+1} = U+1 | U_N = U) \\ &= U\left(1 + \frac{\alpha}{\theta + N}\right) + \frac{\theta}{\theta + N} \end{aligned} \tag{15}$$

The relation of (14) holds by taking the expectation of (15) over U. Q.E.D.

We can similarly calculate the factorial moment:

$$\mathrm{E}\left(\prod_{j=1}^{N} S_j^{(r_j)}\right) = \sum_{U=1}^{N} \frac{N^{(R)}\theta^{[U:\alpha]}}{(\theta+N-1)^{(R)}\theta^{[U-r:\alpha]}} \prod_{j=1}^{N}\left(\frac{(1-\alpha)^{[j-1]}}{j!}\right)^{r_j} P(U_{N-R}=U-r) \tag{16}$$

where $r=\sum_{j=1}^{N} r_j$, $R=\sum_{j=1}^{N} jr_j$ and $n^{(R)}=n(n-1)\cdots(n-R+1)$. The higher moments of U_N are evaluated through recurrence relations like (14). For example

$$\mathrm{E}(U_{N+1}^2)=\frac{N+\theta+2\alpha}{\theta+N}\mathrm{E}(U_N^2)+\frac{2\theta+\alpha}{\theta+N}\mathrm{E}(U_N)+\frac{\theta}{\theta+N}$$

In particular we obtain

$$\mathrm{E}(S_1)=\frac{N\theta+N\alpha\mathrm{E}(U_{N-1})}{\theta+N-1} \tag{17}$$

$$\mathrm{E}(S_2)=\mathrm{E}\left[\frac{N^{(2)}(\theta+\alpha U_{N-2})}{(\theta+N-1)^{(2)}}\right]\frac{1-\alpha}{2} \tag{18}$$

$$\mathrm{E}(S_1^{(2)})=\mathrm{E}\left[\frac{N^{(2)}(\theta+\alpha U_{N-2})(\theta+\alpha(U_{N-2}+1))}{(\theta+N-1)^{(2)}}\right] \tag{19}$$

from (11) and (16). These expectations lead to moment estimators discussed in Section 3.

3. Estimation with the Pitman Model

Here we consider the estimation of the disclosure risk under the Pitman model. All the propositions in this section are proved in the Appendix. We denote sample size by n and sample size indices by $s=(s_1,\ldots,s_n)$. The total number of nonempty cells or clusters is $u=\sum_{i=1}^{n} s_i$. Suppose that n individuals are drawn with simple random sampling without replacement.

3.1. The estimation of the parameters

The Pitman model enjoys the property of exchangeability with respect to individuals in a population, assumed in Lemma 1 of Takemura (1999). Accordingly the marginal distribution of sample individuals coincides with the prior distribution of values of n individuals directly drawn from the superpopulation. That is to say,

$$\mathrm{P}(s_1,\ldots,s_n)=n!\frac{\theta^{[u:\alpha]}}{\theta^{[n]}}\prod_{j=1}^{n}\left(\frac{(1-\alpha)^{[j-1]}}{j!}\right)^{s_j}\frac{1}{s_j!} \tag{20}$$

is obtained by replacing N and U in (8) by n and u. We can show the result in another way. Suppose that N objects are partitioned into classes according to a probability distribution p_N. A partition structure (Kingman 1978) is a sequence $p_1, p_2, \ldots$ of distributions wherein, assuming that an object is deleted uniformly at random from the N objects, the partition of the $N-1$ remaining objects is distributed according to p_{N-1}. The Pitman sampling formula is known to have a partition structure, with the result that (20) holds.

We then construct the Maximum Likelihood Estimators (MLE) of θ and α. Let the logarithm of the right hand side of (20) be L. The MLE is the solution of

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{u-1} \frac{1}{\theta + i\alpha} - \sum_{i=1}^{n-1} \frac{1}{\theta + i} = 0$$

and

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{u-1} \frac{i}{\theta + i\alpha} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{j - \alpha} = 0$$

These simultaneous equations can be solved by the Newton-Raphson method using second derivatives:

$$\frac{\partial^2 L}{(\partial \theta)^2} = -\sum_{i=1}^{u-1} \frac{1}{(\theta + i\alpha)^2} + \sum_{i=1}^{n-1} \frac{1}{(\theta + i)^2}$$

$$\frac{\partial^2 L}{(\partial \alpha)^2} = -\sum_{i=1}^{u-1} \frac{i^2}{(\theta + i\alpha)^2} - \sum_{i=2}^{n} s_i \sum_{j=1}^{i-1} \frac{1}{(j - \alpha)^2} < 0 \tag{21}$$

$$\frac{\partial^2 L}{\partial \theta \partial \alpha} = -\sum_{i=1}^{u-1} \frac{i}{(i\alpha + \theta)^2} < 0$$

To solve the ML estimation, we investigate approximate moment estimators for the starting values of the Newton-Raphson procedure. Our moment estimators are

$$\hat{\theta} = \frac{nuc - s_1(n-1)(2u + c)}{2s_1 u + s_1 c - nc} \tag{22}$$

$$\hat{\alpha} = \frac{\hat{\theta}(s_1 - n) + (n-1)s_1}{nu} \tag{23}$$

where $c = s_1\,(s_1 - 1)/s_2$. The derivation is given in the Appendix. In six of seven cases in Section 4, except for Case 1, these estimators gave convergences in the Newton-Raphson procedure. In Case 1 the author reached the solution by random starting value generation.

We show a useful result for such random generation of α and θ; Proposition 1 enables us to restrict the ranges of parameter values that will be explored.

Proposition 1. *Let the ML estimates of the Pitman model be denoted by α^* and θ^*, and let the ML estimate of the Ewens model be denoted by θ_E. Then $\theta^* < \theta_E$ unless $\alpha^* \leq 0$.*

In view of Proposition 1 we can understand why Samuels (1998) objected to the Pitman model. Samuels found that θ_E was too small for his data set and attempted to obtain a larger estimate of the Ewens parameter θ by the Pitman parameter θ^*. Actually Proposition 1 shows that θ^* is smaller than θ_E in the usual cases of positive α^*. However, our experience of applying the Pitman model to various data sets suggests that the flexibility gained by the additional parameter α greatly improves the fitting.

3.2. Risk inference

In the following we discuss some statistics concerning the disclosure risk under the Pitman model. We state three propositions useful for the disclosure problem.

As regards the risk inference, we will evaluate the expectation of the number of population uniques $\mathrm{E}(S_1)$ with the ML estimates of the parameters. However necessary moments given in Section 3 are not in convenient forms to compute. From Theorem 1 we investigate simple forms of the moments.

Proposition 2 *If $\alpha \neq 0$, the expectation of U_N under (8) is reduced to*

$$\mathrm{E}(U_N) = \frac{\theta}{\alpha}\left(\frac{(\theta+\alpha)^{[N]}}{\theta^{[N]}} - 1\right) \tag{24}$$

The result of Yamato and Sibuya (1999) coincides with (24). We can rewrite (24) using the gamma function. Based on the asymptotic property of the gamma function, we find a useful approximation of $\mathrm{E}(U_N)$, which is a special case of Lemma 2 in Yamato and Sibuya (1999). If N is sufficiently large

$$\mathrm{E}(U_N) = \frac{\theta}{\alpha}\left(\frac{\Gamma(\theta+\alpha+N)\Gamma(\theta)}{\Gamma(\theta+N)\Gamma(\theta+\alpha)} - 1\right) \approx \frac{\Gamma(\theta+1)}{\alpha\Gamma(\theta+\alpha)}N^{\alpha} \tag{25}$$

for $\alpha \neq 0$. Our expression of $\mathrm{E}(S_1)$ depending on $\mathrm{E}(U_{N-1})$ then becomes simpler. We obtain

$$\mathrm{E}(S_1) = \frac{N\Gamma(\theta+\alpha+N-1)\Gamma(\theta+1)}{\Gamma(\theta+N)\Gamma(\theta+\alpha)} \approx \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}N^{\alpha} \tag{26}$$

As a result, the evaluation of $\mathrm{E}(S_1)$ is not very hard, once the ML estimates are obtained.

We denote the proportion of population uniques in the sample to sample uniques by p_u. This is often an index of the disclosure risk. Let us denote the sampling ratio by $f = n/N$. It is natural to estimate the proportion by

$$\hat{p}_u = \frac{\hat{S}_1}{s_1} f \tag{27}$$

where $\hat{S}_1$ would be $\mathrm{E}(S_1)$.

The following propositions may have interesting implications regarding the disclosure problem.

Proposition 3 *Suppose that size indices are distributed according to (8). Then, for $\alpha \geq 0$*

$$\lim_{N\to\infty} \frac{\mathrm{E}(S_1)}{\mathrm{E}(U_N)} = \alpha$$

Proposition 3 suggests that the ratio of population uniques to the number of nonzero frequency groups is α, which is smaller than unity; the implication is consistent with the author's experience that the Ewens model (i.e., $\alpha = 0$) tends to underestimate the number of population uniques. This tendency seems to be in line with other authors' experiences. Since the Ewens model is a limiting form of the conditional Poisson-gamma model, these models give similar population unique estimates, as can be seen in Section 4. It is suggested that the poor performance of the Poisson-gamma model and related models, including the logarithmic series distribution, occurs when population uniques constitute no negligible proportion of the population. In other words these models might be suitable only for safe data sets.

Based on Proposition 3, we propose a simple estimator of α:

$$\tilde{\alpha} = \frac{s_1}{u} \tag{28}$$

We could replace the previous moment estimator (23) by (28); see Table 8 of Section 4.

Proposition 4 *Let $n/N = f$ be fixed. If we assume (8) and (20) then*

$$\lim_{N \to \infty} \frac{\mathrm{E}(S_1)}{\mathrm{E}(s_1)} f = f^{1-\alpha} \tag{29}$$

The left-hand side of (29) corresponds to p_u in (27). Combining the simple estimate of (28), we can roughly evaluate the risk of a data set by $f^{1-\tilde{\alpha}}$, where the sampling ratio f is known. This simple procedure is useful because the data editing for anonymization requires repeated trial and error.

4. An Application to Japanese Labor Force Survey Data

In this section we examine performances of the Pitman model and other superpopulation models by applying them to real data. The data are given by Takemura (1998), from the Japanese labor force survey. First we consider how to compare performances of superpopulation models. Second, we discuss the application results.

4.1. The methodology of comparison

Among the superpopulation models considered above, no model is universally the best on logical grounds. Rather we should evaluate the appropriateness of each model to a given data set. Our risk evaluation will then proceed as follows: (a) fix a group of models, (b) measure the difference between each model in the group and the given data set, and (c) adopt the estimation given by the best model for the data set. The problem is how to measure the goodness of fit of a model to data. In the following we discuss some criteria for the measurement.

An χ^2 type statistic like

$$\chi^2 = \sum_{i=1} \frac{(s_i - \mathrm{E}(s_i))^2}{\mathrm{E}(s_i)} \tag{30}$$

is conventionally used to evaluate the goodness of stochastic abundance model fitting. If $(s_1, \ldots, s_n)$ is multinomially distributed given u, then (30) is the classical χ^2 test. The symmetrical model description in terms of independent F_j's can be converted in terms of S_i's, where

$$\mathrm{P}(S_0, \ldots) = \binom{K}{S_0\, S_1\, \ldots} \prod_{i=0}^{\infty} \mathrm{P}(F = i)^{s_i} \tag{31}$$

is in the form of multinomial distribution. Since the marginal distribution of the multinomial distribution becomes multinomial, the χ^2 type statistic might be suitable for the Poisson-gamma model and the Poisson-lognormal model. However, the assumption seems to be inappropriate for the other superpopulation models. As described in Section 7.2 of

Engen (1978), we can only use (30) to ''form a picture of the similarity'' between s_i's and $\mathrm{E}(s_i)$'s. In the disclosure context, Zayatz (1991) used the Kolmogorov-Smirnov goodness of fit test for the Poisson-gamma model and found a significant lack of fit at the .01 level. Skinner and Holmes (1993) calculated (30) and likelihood ratio statistics for the logarithmic series distribution and the Poisson-lognormal model. These ideas are based on the theory of testing hypotheses, and we can only, at best, tell whether the model assumption is acceptable or not. In other words these statistics are not comparable between different models. They are especially inappropriate for models with different numbers of parameters.

There is also a problem of truncation. Chen and Keller-McNulty (1998) recognized that the size indices of microdata tended to have a heavy upper tail (i.e., the largest i on which s_i is not zero is often large). They proposed fitting a model mainly for small cells (i.e., s_1, s_2 and so on) and disposing the information of the tail. Skinner and Holmes (1993) recommended censoring such a tail. They stated that a truncated lognormal distribution fitted well to microdata. Similar truncation is popular for lognormal fitting in statistical ecology, although its objective seems to be data description rather than parameter estimation. Many people consider that the frequency distribution of small cell sizes may be more important than the upper tail. However, it is not valid to measure the fit by only small cells. For instance, let us consider a fixed population that has size indices $S_1, \ldots, S_N$. From this population, we choose n individuals by simple random sampling without replacement. It was shown by Greenberg and Zayatz (1992) that

$$\mathrm{E}(s_j) = \sum_{i=1}^{N} S_i \binom{i}{j} \binom{N-i}{n-j} \Big/ \binom{N}{n}, \qquad j = 1, \ldots, n \tag{32}$$

Thus it is very likely that even small cells in the sample are composed of small and large cells in the population. It might seem that the information of $s_2, s_3, \ldots$ does not contain information on population uniques, because a population unique is also unique in the sample. However, through such dependency as (32), correct estimation of whole size indices is important even for the estimation of population uniques. In any case it is better to utilize the whole information of the sample.

A possible solution is to use the Akaike Information Criterion (AIC), which is a standard tool for model selection. Let the number of parameters in a model be λ. Let the log likelihood of the model maximized with respect to the parameters be denoted by L. The AIC selects the model that has the lowest $A = -2L + 2\lambda$. See Atkinson (1980) or Konishi and Kitagawa (1996), for example.

Hereafter our model selection is based on the AIC. We next consider the calculation

Table 1. The relationship between a model and its sampling distribution

Population model	Sampling distribution
Poisson-gamma: (1)	Dirichlet-multinomial: (34)
Poisson-lognormal: (2)	Approximate Poisson-lognormal: (35)
Dirichlet-multinomial: (3)	Dirichlet-multinomial: (34)
Logarithmic-series: (4)	Ewens: (33)
Ewens: (5)	Ewens: (33)
Pitman: (8)	Pitman: (20)

of AIC value A for the models discussed in Section 1.1. The likelihood depends on the sampling mechanism; we assume simple random sampling without replacement. In the following we list the sampling distribution of each model. The results are summarized in Table 1.

The sampling distribution of the Pitman model was shown in (20). Similarly, the sampling distribution of the Ewens model is again the Ewens model:

$$\mathrm{P}(s_1, \ldots, s_n) = \frac{\theta^u}{\theta^{[n]}} \frac{n!}{\prod_{i=1}^{n} i^{s_i} s_i!} \tag{33}$$

Hoshino and Takemura (1998) show that the sampling distribution of the Fisher logarithmic series model is the Ewens model (33). See Sibuya (1991) or Hoshino and Takemura (1998) for the parameter estimation of the Ewens model.

According to Takemura (1999), the sampling distribution $\mathrm{P}(s_0, \ldots | n)$ of the Poisson-gamma model given the sample size n becomes the Dirichlet-multinomial model:

$$\mathrm{P}(s_0, \ldots, s_n) = \frac{n! K! \Gamma(K\gamma)}{\Gamma(K\gamma + n)} \prod_{i=0}^{n} \left(\frac{\Gamma(\gamma + i)}{\Gamma(\gamma) i!} \right)^{s_i} \frac{1}{s_i!} \tag{34}$$

Also, the sampling distribution of the Dirichlet-multinomial model is (34).

Under simple random sampling without replacement, the sampling distribution $\mathrm{P}(s_0, \ldots | n)$ of the Poisson-lognormal model is hard to manipulate. Therefore for this model we assume the Bernoulli sampling (Särndal et al. 1992) in which each individual is drawn as if a coin with some success probability comes up in heads. This scheme is a convenient approximation to simple random sampling without replacement, but it is more natural in ecological sampling than simple random sampling without replacement. When the success probability is n/N_0, we obtain the sampling distribution $\mathrm{P}(s_0, \ldots)$ of the form (31) replacing N_0 by n. Another approximation we use is the normal approximation of the sample size distribution. The variance of the sample size becomes $T = K(\exp(M + V/2) + \exp(2M + 2V) - \exp(2M + V))$, and the expected sample size is set to n. Therefore we set the probability of the sample size being n as $1/\sqrt{2\pi T}$. Consequently the conditional Poisson-lognormal model $\mathrm{P}(s_0, \ldots)/\mathrm{P}(\sum i s_i = n)$ approximated by

$$\mathrm{P}(s_0, \ldots | n) = \binom{K}{s_0 \cdots s_n} \prod_{i=0}^{n} \left\{ \frac{1}{i! \sqrt{2\pi V}} \int_0^\infty \lambda^{i-1} \exp(-\lambda - (\log \lambda - M)^2 / 2V) d\lambda \right\}^{s_i} \sqrt{2\pi T} \tag{35}$$

where $M = \log n - \log K - V/2$. We need numerical integration to evaluate the model. A transformation suited to the Hermitian integration is discussed in Aitchison and Ho (1989). The author programmed the numerical integration with the GNU C compiler, checking the results against Grundy (1951).

In short we can compare all the models presented in Section 1.1, by calculating AIC values for the Ewens model (33), the Dirichlet-multinomial model (34), the Pitman model (20) and the approximate Poisson-lognormal model (35).

4.2. An application

Now we sketch the data to be assessed. The purpose of the labor force survey is to elucidate the current state of employment and unemployment in Japan. The data at hand was collected in January 1995. The population of the survey is composed of all individuals 15 years old and over usually residing in 47 prefectures of Japan. However, the data at hand *consists of* individuals only in nine prefectures near Tokyo. The sample size n is 27,230. The corresponding population size N is about 35.85 million. For simplicity we assume that individuals are drawn simply at random without replacement, although the actual sampling scheme is more complicated. Seven different combinations of ''global recoding'' and ''global suppression'' are applied to the data. See Willenborg and de Waal (1996) for these techniques of anonymization. The size indices are enumerated with respect to the categorization of nine (Cases 1–2), eight (Cases 3–6) and seven (Case 7) variables. These variables are geographical codes, classified number of individuals in the household, relationship to head of household, sex, age, and marital status. Table 2 provides more information on the categorization. The results of our model fitting are summarized in Tables 3 to 6. The estimated p_u is given by (27), where $\hat{S}_1$ equals the estimated $E(S_1)$.

The Pitman model highly dominates in all the cases, and the Poisson-lognormal model

Table 2. The number of categories for each variable (Cases 1–7 of Section 4)

	Variable	Case 1	Case 2	Case 3
(A)	Prefecture code	(9)	(9)	(9)
(B)	– zone code	824	824	824
(C)	Individuals 15 years old and over	8	3	3
(D)	– under 15 years (male)	6	3	2
(E)	– under 15 years (female)	5	3	(D)
(F)	Relationship to the head	12	5	5
(G)	Sex	2	2	2
(H)	Age	100	20	10
(I)	Marital status	4	4	4
	K (Cell total)	1,898,496,000	17,798,400	1,977,600

	Variable	Case 4	Case 5	Case 6	Case 7
(A)	Prefecture code	9	9	9	9
(B)	– zone code	x	x	x	x
(C)	Individuals 15 years old & over	8	8	8	8
(D)	– under 15 years (male)	6	6	4	6
(E)	– under 15 years (female)	5	5	4	(D)
(F)	Relationship to the head	12	12	12	12
(G)	Sex	2	2	2	2
(H)	Age	100	20	20	20
(I)	Marital status	4	4	4	4
	K (Cell total)	20,736,000	4,147,200	2,211,840	829,440

NOTE: ''x'' indicates that the variable is suppressed. The underlining indicates that the number of categories is smaller than that of Case 1. ''(D)'' indicates that the information on the variable is represented by the variable (D).

Table 3. Cases 1-2 of Section 4

	Case 1	Case 2
Total cell number (K)	1,898,496,000	17,798,400
Total nonempty cell number (u)	25,923	21,851
Sample uniques (s_1)	25,046	18,275
Maximum cell size	28	28
Ewens parameter θ by MLE	280628.969879	52004.115657
Log likelihood (AIC)	−518.9 (1039.7)	−245.3 (492.6)
Estimated population uniques E(S_1)	278449.3	51928.8
Estimated p_u	0.84%	0.22%
Dirichlet-multi parameter γ by MLE	0.000148	0.002928
Log likelihood (AIC)	−518.9 (1039.8)	−246.3 (494.6)
Estimated population uniques E(S_1)	278244.1	51044.0
Estimated p_u	0.84%	0.21%
Pitman parameters α, θ by MLE	0.917448, 16389.753923	0.520587, 21297.598824
Log likelihood (AIC)	−111.8 (227.6)	−100.9 (205.8)
Estimated population uniques E(S_1)	19000174.4	1017904.0
Estimated p_u	57.6%	4.23%
Poisson-lognormal parameter V by MLE	10.530755	8.524957
Log likelihood (AIC)	−547296.6 (1094595.2)	−5750.6 (11503.1)
Estimated population uniques E(S_1)	11950655.9	1773952.1
Estimated p_u	36.2%	7.37%

Table 4. Cases 3–4 of Section 4

	Case 3	Case 4
Total cell number (K)	1,977,600	20,736,000
Total nonempty cell number (u)	18,221	12,390
Sample uniques (s_1)	12,919	8049
Maximum cell size	41	54
Ewens parameter θ by MLE	24249.278863	8804.206385
Log likelihood (AIC)	−142.0 (286.1)	−686.0 (1374.0)
Estimated population uniques E(S_1)	24232.9	8802.0
Estimated p_u	0.14%	0.083%
Dirichlet-multi parameter γ by MLE	0.012424	0.000425
Log likelihood (AIC)	−143.8 (289.6)	−686.7 (1375.4)
Estimated population uniques E(S_1)	22427.7	8769.8
Estimated p_u	0.13%	0.083%
Pitman parameters α, θ by MLE	0.140768, 19948.932049	0.501239, 2585.173765
Log likelihood (AIC)	−131.4 (266.8)	−101.9 (207.7)
Estimated population uniques E(S_1)	57260.1	308054.4
Estimated p_u	0.34%	2.91%
Poisson-lognormal parameter V by MLE	5.166813	14.268244
Log likelihood (AIC)	−4235.1 (8472.2)	−14134.2 (28270.4)
Estimated population uniques E(S_1)	357874.6	495826.6
Estimated p_u	2.10%	4.68%

Table 5. Cases 5–6 of Section 4

	Case 5	Case 6
Total cell number (K)	4,147,200	2,211,840
Total nonempty cell number (u)	6,657	6,653
Sample uniques (s_1)	3,813	3,805
Maximum cell size	154	154
Ewens parameter θ by MLE	2813.718472	2810.978767
Log likelihood (AIC)	−997.1 (1996.2)	−993.1 (1988.2)
Estimated population uniques E(S_1)	2813.5	2810.8
Estimated p_u	0.056%	0.056%
Dirichlet-multi parameter γ by MLE	0.000679	0.001271
Log likelihood (AIC)	−998.6 (1999.3)	−996.0 (1993.9)
Estimated population uniques E(S_1)	2795.9	2778.0
Estimated p_u	0.056%	0.055%
Pitman parameters α, θ by MLE	0.505272, 523.377001	0.504301, 525.742679
Log likelihood (AIC)	−219.7 (443.3)	−219.7 (443.5)
Estimated population uniques E(S_1)	145294.2	144053.2
Estimated p_u	2.89%	2.88%
Poisson-lognormal parameter V by MLE	14.370145	13.053184
Log-likelihood (AIC)	−13291.8 (26585.5)	−10398.8 (20799.7)
Estimated population uniques E(S_1)	320562.6	199167.2
Estimated p_u	6.39%	3.98%

Table 6. Case 7 of Section 4

	Case 7
Total cell number (K)	829,440
Total nonempty cell number (u)	5,682
Sample uniques (s_1)	2,974
Maximum cell size	154
Ewens parameter θ by MLE	2188.670938
Log likelihood (AIC)	−759.9 (1521.7)
Estimated population uniques E(S_1)	2188.5
Estimated p_u	0.056%
Dirichlet-multi parameter γ by MLE	0.002646
Log likelihood (AIC)	−764.8 (1531.6)
Estimated population uniques E(S_1)	2138.9
Estimated p_u	0.055%
Pitman parameters α, θ by MLE	0.443278, 524.588977
Log likelihood (AIC)	−228.2 (460.3)
Estimated population uniques E(S_1)	72949.3
Estimated p_u	1.86%
Poisson-lognormal parameter V by MLE	9.209263
Log likelihood (AIC)	−10761.7 (21525.4)
Estimated population uniques E(S_1)	108974.3
Estimated p_u	2.78%

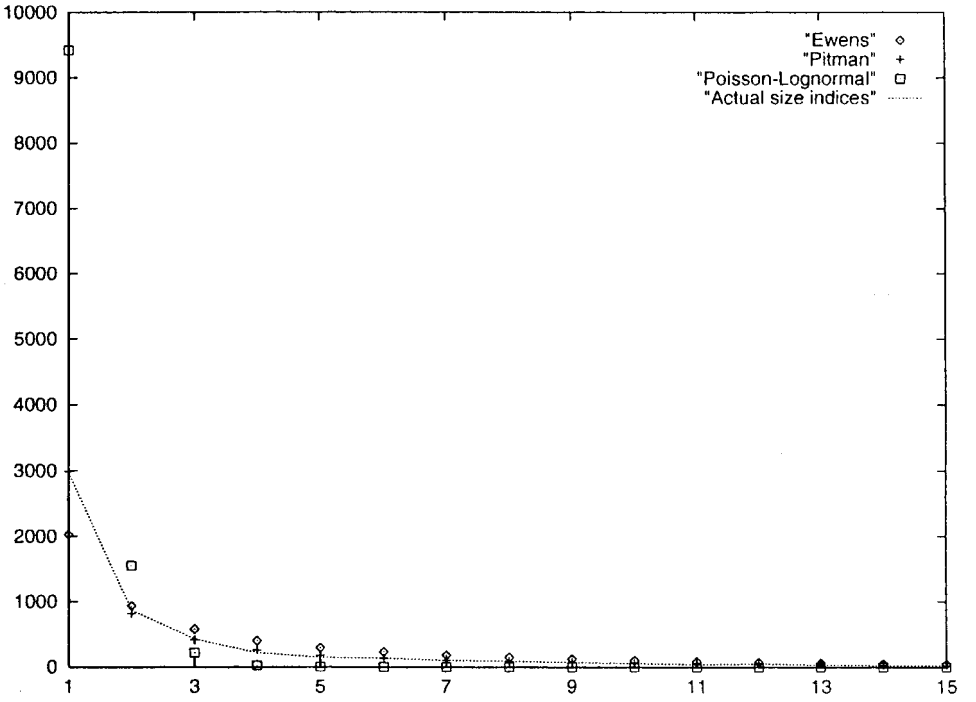

Fig. 1. Fits of $E(s_i)$'s under ML estimates (Case 7, Section 4)

shows the least performance. Engen (1978) and Skinner and Holmes (1993) reported relatively good fits of the Poisson-lognormal model based on the χ^2 type statistic in (30). It might be the case that for the Poisson-lognormal model the maximization of the marginal likelihood on $(s_1, \ldots, s_n)$ gives a different estimate. Figure 1 illustrates the fits of the Ewens model, the Pitman model and the Poisson-lognormal model in Case 7. The vertical axis corresponds to $E(s_i)$'s under the ML estimates of the parameters, and the horizontal axis corresponds to $i = 1, \ldots, 15$. The actual sample size indices are plotted in the same scale. Under the Poisson-lognormal model, $E(s_1)$ shows enormous overshoot. This may indicate that the inclusion of zero frequency groups causes the lowest fit of the Poisson-lognormal model. It should be noted that the Pitman model ignores the restriction that K is finite. Thus for fairer comparison, we explore $K = K^*$ in which the Poisson-lognormal model attains the smallest AIC value; it is a kind of marginal fitting. Chen and Keller-McNulty (1998) give a similar kind of marginal fitting for the Poisson-gamma model. Our results are provided in Table 7. The Pitman model still dominates except for Case 3. We accordingly observe the strong support of the Pitman model.

Table 7. Poisson-lognormal model fits with optimized K (Section 4)

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
K^*	657,385	198,613	71,065	36,594	10,258	10,254	7,883
Log likelihood	−635.2	−157.5	−122.5	−864.9	−2153.4	−2148.4	−2103.6
AIC	1272.4	316.9	246.9	1731.9	4308.7	4298.7	4209.3
$E(S_1)$ with K^*	3894.5	431.9	1.5	5.3	0.0	0.0	0.0
$E(S_1)$ with K	33902689.2	4076872.5	104255.6	4273132.0	480176.7	147338.8	8979.4

Table 8. The simple estimates of the Pitman parameter α (Section 4)

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
s_1/u	0.97	0.84	0.71	0.65	0.57	0.57	0.52
α by MLE	0.92	0.52	0.14	0.50	0.50	0.50	0.44

However, we further analyze the results for more detailed evaluation. Case 3 shows some peculiarity when we apply the simple estimator (28) of α. Table 8 lists differences between the MLE and (28); there is a large difference in Case 3 of Table 8. We may be able to regard (28) as a model check. Actually, it seems that the p_u estimated by the Pitman model is small in Case 3. Since there exists no all-purpose estimation procedure, we should probably examine the possibility of an alternative approach.

Let us then return to Table 7. We first realize that K is much larger than K^*, and $\mathrm{E}(S_1)$ with K^* is very small. These facts suggest that the fit of the Poisson-lognormal model has no robustness in withstanding changes of K. The presence of structural zeros, for example caused by the cross-classification of age and marital status, may lead us to accept a claim that true K is smaller than the product of the number of categories in the variables assessed. Although the author considers that structural zeros are also realizations from a superpopulation, let us consider the possibility of such a decrease in K. We may regard K^* as an estimate of the true K, but it seems that K^* is too small; a large underestimate of K implies an underestimate of population uniques. Concerning the risk inference, an underestimate should be more heavily penalized than an overestimate. It is thus not persuasive to believe that K^* equals the true K. Note that arbitrariness cannot be eliminated in determining the true K. It is therefore preferable that the risk inference does not depend on K. However, this independence does not seem to hold in the Poisson-lognormal model. Furthermore Engen (1978) provides an example where the parameter of the Poisson-lognormal model estimated by marginal (excluding zero groups) fitting varies with respect to the size of the sample from the same population. It suggests that the use of K^* by marginal fitting leads to an erroneous estimate of population uniques.

Figure 1 clearly represents a typical tendency of model fitting in the disclosure field. We often observe a large difference between s_1 and s_2. The author considers that a ''shape'' parameter is required to describe this kind of nonsmoothness. In view of the urn model implication, the Pitman parameter α specially adjusts the rate of unique cells. This fact would be the reason why the Pitman model dominates.

5. Discussion

5.1. The Pitman model and the lognormal distribution

Construction 16 of Pitman (1995) provides another derivation of the Pitman sampling formula. In this section we observe that it gives a justification similar to that of the lognormal distribution for the Pitman model. This interpretation may motivate the Pitman model.

The lognormal distribution has long been used to describe various populations of species, savings in households, mineral gains and numerous seemingly unrelated objects. Halmos (1944) gave the following justification of the wide applicability of the lognormal

distribution. Let $\boldsymbol{W} = (W_1, W_2, \ldots)$ be a sequence of random variables, where $0 \leq W_i \leq 1$, $i = 1, 2, \ldots$. Define $\bar{W}_i = 1 - W_i$. Let

$$P_i = \bar{W}_1 \cdots \bar{W}_{i-1} W_i, \qquad i = 1, 2, \ldots \tag{36}$$

Then $\boldsymbol{P} = (P_1, P_2, \ldots)$ constitutes a random classification where the i-th group has proportion P_i. Equation (36) implies that $P_i = (1 - P_1 - P_2 - \cdots - P_{i-1})W_i$. Hence the process allocates the residual. The logarithm of (36) equals

$$\log P_i = \log \bar{W}_1 + \cdots + \log \bar{W}_{i-1} + \log W_i$$

The right-hand side is a sum of random variables. Thus under appropriate regularity conditions the central limit theorem holds, and $\log P_i$ is normally distributed. That is to say, P_i is subject to the lognormal distribution in many cases.

Assume that X_j, $j = 1, \ldots, N$, are independently identically distributed given $\boldsymbol{P}$ with $P(X_j = i|\boldsymbol{P}) = P_i$, $i = 1, 2, \ldots$. Here X_j is the j-th sample from the infinite population of individuals. We can interpret P_i as the long-run relative frequency of the i-th group. The marginal distribution of the frequency $F_i = \sum_{j=1}^{N} I(X_j = i)$ given $\boldsymbol{P}$ is the binomial distribution:

$$P(F_i = y|\boldsymbol{P}) = \binom{N}{y} P_i^y (1 - P_i)^{N-y}, \qquad y = 0, 1, \ldots, N$$

It is well known that the binomial distribution above is approximated by the Poisson distribution with mean $N P_i$. If $\log N P_i = \log N + \log P_i$ is subject to the normal distribution, then the marginal frequency of the i-th group is approximately the Poisson-lognormal.

We now turn to the Pitman model. Let us suppose that W_i of (36) independently possesses the beta distribution with parameters $(1 - \alpha, \theta + i\alpha)$, where $0 \leq \alpha < 1$, $\theta > -\alpha$. Now we can explicitly derive the distribution of the samples $X_1, \ldots, X_N$. According to Pitman (1995), the size indices of the samples are then subject to (8): the Pitman model.

We have observed that the same residual allocation structure induces the Pitman model and the lognormal distribution. The Pitman model may consequently have a motivation for use in the disclosure field, because the Poisson-lognormal model has been used to measure the risk. Yamato et al. (1999) clearly explain the derivation of the Pitman model from the beta distribution. The corresponding derivation of the Ewens model is given in Johnson et al. (1997). The process of (36) with independent W_i is known as the residual allocation model. See Pitman (1996) for a survey.

5.2. *Concluding remarks*

On the Japanese labor force survey data sets, the Pitman model provided the most plausible inference among the existing models. It should also be emphasized that the computation on the Pitman model is not so heavy as the Poisson-lognormal model. Thus it seems that the Pitman model is a promising tool for the disclosure risk assessment. This section appends a few arguments in this regard.

The uniques constitute the lower tail of a distribution. We generally face difficulties in estimation problems concerning a tail. For instance, even the approximation of simple random sampling without replacement by the Bernoulli sampling might considerably

affect the distribution of the lower tail; preferable models are those that employ no approximation in the sampling scheme. The Pitman model, because of its partition structure, is consistent with the sampling scheme in the disclosure field. With a stratified sampling structure, we may apply the Pitman model in each stratum.

However, we should note that the Pitman model ignores the restriction of K. If there is a large difference between the sample size and the population size, then a disregard of K may cause an overestimate of the risk; it is, in the extreme, possible that U becomes larger than the disregarded K. In applying the Pitman model, we should check whether U is too large compared to K.

Appendix

We first derive the moment estimators (22) and (23). For simplicity we use $\mathrm{E}(s_1)$ and $\mathrm{E}(s_1(s_1-1))/\mathrm{E}(s_2)$ to estimate θ and α. We denote the total number of clusters given $n-1$ and $n-2$ by u_{n-1} and u_{n-2}. By (17)

$$\alpha = \frac{(\theta+n-1)\mathrm{E}(s_1) - n\theta}{n\mathrm{E}(u_{n-1})} \tag{37}$$

Referring to (18) and (19), we derive

$$C = \frac{\mathrm{E}(s_1(s_1-1))}{\mathrm{E}(s_2)} = \frac{2\mathrm{E}[(\theta+\alpha u_{n-2})(\theta+\alpha(u_{n-2}+1))]}{(1-\alpha)\mathrm{E}(\theta+\alpha u_{n-2})}$$

$$\approx \frac{2\mathrm{E}(\theta+\alpha(u_{n-2}+1))}{(1-\alpha)}$$

Then

$$\alpha = \frac{C-2\theta}{2\mathrm{E}(u_{n-2})+C+2} \tag{38}$$

From (37) and (38),

$$\theta = \frac{n\mathrm{E}(u_{n-1})C - (n-1)\mathrm{E}(s_1)(2\mathrm{E}(u_{n-2})+C+2)}{(\mathrm{E}(s_1)-n)(2\mathrm{E}(u_{n-2})+C+2)+2n\mathrm{E}(u_{n-1})} \tag{39}$$

Now we give the moment estimator of θ. Ignoring the relation $\mathrm{E}(u_{n-1}) = (\theta+n)/(\theta+n+\alpha)\mathrm{E}(u_n) - \theta/(\theta+n+\alpha)$, we replace $\mathrm{E}(u_{n-1})$ of (39) by u and $\mathrm{E}(u_{n-2})$ by $u-1$, whereby the estimator becomes simpler. Let $C = c = s_1(s_1-1)/s_2$. Substituting $\mathrm{E}(s_1)$ of (39) by s_1, we obtain (22). Equation (23) is a direct consequence of (37).

In the following we show the propositions stated in Section 3.

Proof of Proposition 1. Note that θ_1 is the conditional ML estimate of the Pitman model given $\alpha = 0$. It is widely known that the Ewens distribution belongs to the exponential family. Thus $\partial L/\partial\theta < 0$ for $\theta > \theta_1$ with $\alpha = 0$ (see p. 417 of Lehmann (1991), for example). By (21), $\partial L/\partial\theta < 0$ for $\theta > \theta_1$ with $\alpha > 0$. Consequently, θ^* is never larger than θ_1 if $\alpha^* > 0$. Q.E.D.

Proof of Proposition 2. We derive (24) from (12). The proposition is shown by the relation that

$$
\begin{aligned}
1 + \frac{\alpha \mathrm{E}(U_N)}{\theta} &= 1 + \frac{\alpha}{\theta + N - 1} + \sum_{l=0}^{N-2} \frac{\alpha}{\theta + l} \prod_{j=l+1}^{N-1} \left(\frac{\theta + j + \alpha}{\theta + j} \right) \\
&= \frac{\theta + N - 1 + \alpha}{\theta + N - 1} \left\{ 1 + \sum_{l=0}^{N-3} \frac{\alpha}{\theta + l} \prod_{j=l+1}^{N-2} \left(\frac{\theta + j + \alpha}{\theta + j} \right) + \frac{\alpha}{\theta + N - 2} \right\} \\
&= \frac{(\theta + N - 1 + \alpha)(\theta + N - 2 + \alpha)}{(\theta + N - 1)(\theta + N - 2)} \\
&\quad \times \left\{ 1 + \sum_{l=0}^{N-4} \frac{\alpha}{\theta + l} \prod_{j=l+1}^{N-3} \left(\frac{\theta + j + \alpha}{\theta + j} \right) + \frac{\alpha}{\theta + N - 3} \right\} \\
&\quad \vdots \\
&= \frac{(\theta + \alpha)^{[N]}}{\theta^{[N]}}
\end{aligned}
$$

Q.E.D.

To prove Proposition 3, we need the lemma below.

Lemma 1. *For* $\alpha \geq 0$,

$$\lim_{N \to \infty} \mathrm{E}(U_N) = \infty$$

Proof. With respect to nonnegative α, $\mathrm{E}(U_N)$ is monotonically increasing. Thus it suffices to show that $\mathrm{E}(U_N)$ diverges at $\alpha = 0$. When α equals zero

$$\mathrm{E}(U_N) = \sum_{l=0}^{N-1} \frac{\theta}{\theta + l}$$

from (12), and it is well known that the right-hand side diverges as $N \to \infty$. Q.E.D.

Proof of Proposition 3. From (11) we obtain

$$\frac{\mathrm{E}(S_1)}{\mathrm{E}(U_N)} = \frac{N\alpha}{\theta + N - 1} \frac{\mathrm{E}(U_{N-1})}{\mathrm{E}(U_N)} + \frac{N\theta}{\theta + N - 1} \frac{1}{\mathrm{E}(U_N)}$$

Since

$$\mathrm{E}(U_N) = \mathrm{E}(U_{N-1}) \left(1 + \frac{\alpha}{\theta + N - 1} \right) + \frac{\theta}{\theta + N - 1}$$

from (14), $\mathrm{E}(U_N)/\mathrm{E}(U_{N-1}) \to 1$ as $N \to \infty$. Also $1/\mathrm{E}(U_N) \to 0$ by Lemma 1. We consequently obtain the formula. Q.E.D.

Proof of Proposition 4. By the relation of (26),

$$\frac{\mathrm{E}(S_1)}{\mathrm{E}(s_1)} = \frac{N\Gamma(\theta+\alpha+N-1)\Gamma(\theta+n)}{n\Gamma(\theta+N)\Gamma(\theta+\alpha+n-1)}$$

The formula then holds because of the asymptotic relation of the gamma function.

Q.E.D.

6. References

Aitchison, J. and Ho, C.H. (1989). The Multivariate Poisson-log Normal Distribution. Biometrika, 76, 4, 643–653.

Anscombe, F.J. (1950). Sampling Theory of the Negative Binomial and Logarithmic Series Distributions. Biometrika, 37, 358–382.

Atkinson, A.C. (1980). A Note on the Generalized Information Criterion for Choice of a Model. Biometrika, 67, 2, 413–418.

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Bulmer, M.G. (1974). On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. Biometrics, 30, 101–110.

Chen, G. and Keller-McNulty, S. (1998). Estimation of Identification Disclosure Risk in Microdata. Journal of Official Statistics, 14, 79–95.

Engen, S. (1978). Stochastic Abundance Models. London: Chapman and Hall.

Ewens, W.J. (1972). The Sampling Theory of Selectively Neutral Alleles. Theoretical Population Biology, 3, 87–112, and erratum, p. 376.

Ewens, W.J. (1990). Population Genetics Theory – the Past and the Future. In Mathematical and Statistical Development of Evolutionary Theory, ed., S. Lessard 177–227. Dordrecht: Kluwer.

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.

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

Greenberg, B.V. and Zayatz, L.V. (1992). Strategies for Measuring Risk in Public Use Microdata File. Statistica Neerlandica, 46, 33–48.

Grundy, P.M. (1951). The Expected Frequencies in a Sample of an Animal Population in Which the Abundances of Species Are Log-normally Distributed. Biometrika, 38, 427–434.

Halmos, P.R. (1944). Random Alms. Annals of Mathematical Statistics, 15, 182–189.

Hoshino, N. and Takemura, A. (1998). Relationship Between Logarithmic Series Model and Other Superpopulation Models Useful for Microdata Disclosure Risk Assessment. Journal of the Japan Statistical Society, 28, 2, 125–134.

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

Johnson, N.L., Kotz, S., and Balakrishnan, N. (1997). Discrete Multivariate Distributions, Chap. 41, New York: Wiley.

Kingman, J.F. (1978). Random Partitions in Population Genetics. Proceedings of the Royal Society of London, A, 361, 1–20.

Konishi, S. and Kitagawa, G. (1996). Generalised Information Criteria in Model Selection. Biometrika, 83, 4, 875–890.

Lehmann, E.L. (1991). Theory of Point Estimation. California: Wadsworth.

Pitman, J. (1995). Exchangeable and Partially Exchangeable Random Partitions. Probability Theory and Related Fields, 102, 145–158.

Pitman, J. (1996). Random Discrete Distributions Invariant Under Size-Biased Permutation. Advances in Applied Probability, 28, 525–539.

Pitman, J. and Yor, M. (1997). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. Annals of Probability, 25, 855–900.

Samuels, S.M. (1998). A Bayesian Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment. Journal of Official Statistics, 14, 373–383.

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

Sibuya, M. (1991). A Cluster-Number Distribution and its Application to the Analysis of Homonyms. Japanese Journal of Applied Statistics, 20, 139–153 (in Japanese).

Sibuya, M. (1993). A Random-Clustering Process. Annals of Institute of Statistical Mathematics, 45, 459–465.

Skinner, C.J. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. Statistica Neerlandica, 46, 21–32.

Skinner, C.J. and Holmes, D.J. (1993). Modelling Population Uniqueness. In Proceedings of the International Seminar on Statistical Confidentiality, Dublin.

Takemura, A. (1998). Size Indices Data from the Japanese Labor Force Survey. Report of the research supported by the Grant No. 09206102 for Scientific Research of the Ministry of Education, 95–104 (In Japanese).

Takemura, A. (1999). Some Superpopulation Models for Estimating the Number of Population Uniques. In Statistical Data Protection – Proceedings of the Conference, Lisbon, 25 to 27 March 1998–1999 edition, 45–58, Office for Official Publications of the European Communities, Luxembourg.

Watterson, G.A. (1973). Models for the Logarithmic-Species Abundance Distributions. Theoretical Population Biology, 6, 217–250.

Willenborg, L. and de Waal, T. (1996). Statistical Disclosure Control in Practice, Lecture Notes in Statistics 111, New York: Springer.

Yamato, H. and Sibuya, M. (1999). Moments and Asymptotic Distribution of Number of Components of Partition Associated with Pitman Sampling Formula. (submitted).

Yamato, H., Sibuya, M., and Nomachi, T. (1999). Ordered Sample from Two-Parameter GEM Distribution. (submitted).

Zayatz, L.V. (1991). Estimation of the Percent of Unique Population Elements in a Microdata File Using the Sample. Statistical Research Division Report RR-91/08, U.S. Census Bureau.

Received June 2000
Revised May 2001