The Quasi-multinomial Distribution as a Tool for Disclosure Risk Assessment

Nobuaki Hoshino[*]
Faculty of Economics, Kanazawa University

This version: September 2008

Abstract

To model diverse populations, the present article proposes a family of distributions that are closed under recoding and suppression. This family is supported by an empirical fact known as Zipf's law and suggests an upper bound of disclosure risk. The quasi-multinomial distribution is an instance from this family and studied in particular for the assessment of disclosure risk. Also reported is an example in which the quasi-multinomial distribution fits better than existing models.

Keywords: Compound Poisson, Privacy, Size Indices, Superpopulation, Uniqueness

1 Introduction

The precondition of disseminating microdata is to assess correctly the risk of disclosing surveyees' privacy. However, disclosure risk is difficult to evaluate when a sampling fraction is not large. The reason of this difficulty is the lack of information about a population. Superpopulation models thus compensate auxiliary knowledge about a population. The most crucial point here is the proximity of a model to the reality.

Because of the broad variety of real populations, a single model would be unable to approximate every case. Hence we should employ a set of models, among which the best fitted model is to provide plausible assessment given data. We therefore need construct a wide set of models, but little argument seems to exist about the method of construction.

The present article constructs a family of superpopulation models that are closed under anonymization techniques. In other words, an anonymized model belongs again to the same family. This reproductive property is desirable since the degree of anonymization is determined after trial-and-errors. Throughout the process, we should use the same ruler.

To introduce this family, let us begin with considering the general structure of our field's common practice. Suppose that microdata are published after anonymization. A microdata set consists of records with information of fields, and anonymization techniques are applied to selected records and fields. A basic technique of anonymization is *recoding*, which coarsens the

[*] *Address for correspondence* : Nobuaki Hoshino, Faculty of Economics, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp

information of a field by reducing the number of categories of the field. Another basic technique is *suppression*, by which a field is concealed.

First we consider applying the same combination of these two techniques *globally* to all records, obedient to prevailing practice. *Global suppression* is not customary wording, but this term here indicates that a field is unpublished. *Local* anonymization techniques, by which a part of records is modified, will be discussed together with perturbative ones in Section 6. Readers unfamiliar with anonymization concepts should consult with Willenborg and de Waal (1996, 2001).

Under the combination of global recoding and suppression, the risk assessment of microdata reduces to that of a contingency table. The global property in this case implies that all individuals (records) are classified by the same categorization of fields. Hence a single contingency table losslessly summarizes the information of records. The present article thus considers modeling of a contingency table.

Global suppression is nothing but the collapse of cells of a contingency table. That is, global suppression reduces the dimension of a table by merging cells that have the same values of variables except for the suppressed variables. Recoding is similarly equivalent to the collapse of cells. Therefore our model should be treatable on collapsing or merging cells.

Let $\mathbb{N}_0$ and $\mathbb{N}$ be the set of nonnegative integers and the set of positive integers respectively. For $n \in \mathbb{N}$, $[n] := \{1, 2, \ldots, n\}$. We use the following notation for a contingency table. The number of key variables (for the identification of an individual) determines the dimension of a table, and the product of the number of categories of key variables is the total number of cells, which is denoted by J. We append a one dimensional index $j, j \in [J]$, to each cell. A natural index of a cell may be multidimensional, but the location of a cell in a multidimensional space can be a function of j. This information consists in the covariates of a cell. The frequency of the j-th cell is denoted by F_j, and

$$\boldsymbol{F}_J := (F_1, F_2, \ldots, F_J), \quad J \in \mathbb{N}.$$

The total number of individuals is

$$N := \sum_{j=1}^{J} F_j.$$

A typical risk measure of a file is the number of unique cells, each of which contains only one individual. More generally the number of small cells is important risk information, as Greenberg and Zayatz (1992) pointed out. We define frequencies of frequencies (Good (1953)) or size indices (Sibuya (1993)) as

$$S_i := \sum_{j=1}^{J} I(F_j = i), \tag{1}$$

where $I(\cdot)$ is an indicator. The number of unique cells is then expressed by S_1. For a vector, we write

$$\boldsymbol{S}_n := (S_1, S_2, \ldots, S_n), \quad n \in \mathbb{N}.$$

To discuss risk measures precisely, we have to discriminate between a population and samples. For example, unique cells in a population are called population uniques, different from sample uniques. A sample unique that is also a population unique is called a special unique; see Dale and Elliot (2001) for example. Several authors use the proportion of the number of special uniques to the number of sample uniques as a risk measure. The point estimate of the number of special uniques can be a sampling fraction times the number of population uniques. Similarly the weighted sum of population size indices has been proposed as a risk measure. For example, Bethlehem et al. (1990) use the inverse of

$$\sum_{i=1}^{N}\left(\frac{i}{N}\right)^2 S_i,$$

and Greenberg and Zayatz (1992) propose the use of

$$-\sum_{i=1}^{N}\log\left(\frac{i}{N}\right)\frac{i}{N}S_i.$$

Therefore, except for a census, the estimation of population size indices is important.

The unique unbiased estimator of a size index was given by Engen (1978, Theorem 2.1) under simple random sampling without replacement for a fixed population. However, this estimator is useless because of its vast variance; see Shlosser (1981) for one explanation. Hence we assume a distribution of size indices or a superpopulation to supply auxiliary information.

Concerning the risk measure of a record, we need evaluate a function of F_j, where j is the index of a cell to which an evaluated record belongs. A population's F_j is also difficult to estimate when a sampling fraction is low. Therefore we still need a superpopulation.

The present paper regards both a population and samples as realizations of the same superpopulation. Namely a population consists of new draws from a fixed superpopulation. Then we assess the risk by an empirical Bayes method: the parameters of a superpopulation are estimated, based on which we take the expectation of a risk measure.

For example, Franconi and Polettini (2004) measured the risk of a record by the posterior mean of $1/F_j$ given the sample frequency of the jth cell. However, instead of this Bayesian approach, our assessment progresses in the following way. First we estimate the parameter θ of a model based on samples, which is denoted by $\hat{\theta}$. Let the corresponding population size be n^*. Then we evaluate the risk as

$$\mathrm{E}(1/F_j|F_j \geq 1, N = n^*; \hat{\theta}), \qquad (2)$$

given that the j-th cell is a sample unique. Since a population frequency is never less than a sample frequency, (2) is conditioned on $F_j \geq 1$. As regards population uniques, simply we evaluate $\mathrm{E}(S_1|N = n^*; \hat{\theta})$.

It is noteworthy that F_j and S_i are not necessarily of a population. They express random variables, and

$$\boldsymbol{f}_J := (f_1, f_2, \ldots, f_J), \quad \boldsymbol{s}_n := (s_1, s_2, \ldots, s_n)$$

denote just realized values.

In the next section, we construct a family of the distributions of $\boldsymbol{F}_J$ that are closed under global recoding and suppression. This family unifies the existing context of superpopulation models commenced by Bethlehem et al. (1990). Moreover a new method of risk assessment results from an instance of this family: the Quasi-Multinomial (QM) distribution proposed by Consul and Mittal (1975, 1977). We formally introduce the QM distribution in Section 3. A special treatable case of the QM distribution is considered in Section 4. Risk assessment by the QM distribution is demonstrated in Section 5. We clarify the scope of the proposed family for general anonymization practices in Section 6. All the proofs of Theorems are given in Appendix.

2 A family of distributions consistent with anonymization

In this section, we propose a family of distributions for modeling a contingency table. This family contains the multinomial distribution and the Dirichlet-multinomial mixture, which is often used in statistical disclosure control as well; see Takemura (1999), Omori (1999) or Forster and Webb (2007) for various uses. The QM distribution is a member of this family and thus potentially useful in diverse contexts.

The most basic way to model the frequencies of a contingency table would be to suppose that each F_j is independently Poisson distributed with mean θ_j. Then without restriction, the number of the parameters is J, which tends to be very large in our field. Hence we would like to introduce some restriction. For example, a log-linear model determines the number of parameters so that the variance of F_j equals θ_j, which is explained by covariates. However, covariates may be so uninformative that the variation of θ_j is not sufficiently described and F_j overdisperses. This is common because we have to deal with a sparse contingency table. Another special reason of our field is that covariates may be perturbed, which decreases the information. Hence the most basic modeling should be modified to manage overdispersion.

One method is to change the distribution of F_j. For example, the negative binomial distribution is widely used for this purpose. Nevertheless other distributions should be usable, and we are interested in the family of usable distributions. To find this family, we consider the most overdispersed case: all F_j are independent and identically distributed. If a distribution can fit this case, it should also fit a less overdispersed case where the means of cells are not identical.

The most overdispersed model depends on no covariate. It is thus better to supply information other than covariates. The present article employs the following empirical fact. In numerous fields such as linguistics or statistical ecology, size indices are often log-convex:

$$\frac{S_{i+1}}{S_i} \geq \frac{S_i}{S_{i-1}}, \quad i \in \mathbb{N}.$$

See Figure 1 in Section 5 for one example in our field. We can regard this tendency as a version of Zipf's law; see Zipf (1949) or Mandelbrot (1983). Therefore a model should be consistent with the log-convexity. In Proposition 1, cited from Hoshino (2004), the log-convexity naturally restricts the distribution of F_j.

Proposition 1 *Let $F_j, j \in [J]$, be independent and identically distributed. Suppose that* $\mathrm{P}(F_j = 0) \neq 0$ *and* $\mathrm{P}(F_j = 1) \neq 0$*. If the expectations of size indices are log-convex or*

$$\frac{\mathrm{E}(S_{i+1})}{\mathrm{E}(S_i)} \geq \frac{\mathrm{E}(S_i)}{\mathrm{E}(S_{i-1})}, \quad i \in \mathbb{N}, \tag{3}$$

then F_j is compound Poisson distributed.

Therefore, the distribution of F_j should be compound Poisson, which is defined by the following probability generating function (pgf):

$$G_j(z) := \exp(\theta_j(g(z) - 1)), \quad 0 \leq \theta_j < \infty, \tag{4}$$

where

$$g(z) = \sum_{i=1}^{\infty} q_i z^i \tag{5}$$

is another pgf of a distribution over positive integers: q_i is the probability of $i, i \in \mathbb{N}$. We assume that these pgfs are convergent for z in a neighborhood of zero and the distribution $\{q_i\}$ is proper. The parameter θ_j is proportional to $\mathrm{E}(F_j)$ and allowed to vary among cells.

The quintessence of (4) is the negative binomial distribution, where $g(z)$ is of the logarithmic series distribution. The negative binomial equals the Poisson distribution mixed with the gamma distribution, but the mixed Poisson is a different concept from the compound Poisson. See Steutel and van Harn (2004, p.368) for a relationship between these concepts. A compound Poisson distribution is also called a Poisson-stopped-sum distribution; see Johnson et al. (1993, p.351). Sometimes (4) is designated the generalized Poisson distribution because it reduces to the Poisson distribution when $g(z) = z$. A compound Poisson distribution overdisperses, as stated in Johnson et al. (1993, p.354).

Consequently, the joint distribution of $\boldsymbol{F}_J$ should be the product of independent compound Poisson distributions. Moreover, conditioning on N is more natural than dealing with random N, since the total frequency of a population is usually known in risk assessment. Therefore we consider the joint conditional distribution:

$$\mathrm{P}(\boldsymbol{F}_J = \boldsymbol{f}_J | N = n). \tag{6}$$

Another defense of this conditioning is the accuracy of risk inference, which will be formally stated by Theorem 2. Now we define our family of interest.

Definition 1 *Suppose that $F_j, j \in [J]$, is independently compound Poisson distributed as* (4)*, where $\sum_{j=1}^J \theta_j > 0$. We then call the conditional distribution* (6) *a Conditional Compound Poisson (CCP) distribution generated by the distribution of $g(z)$. The parameter*

$$\pi_j := \frac{\theta_j}{\sum_{j=1}^J \theta_j}, \quad j \in [J], \tag{7}$$

is called a cell probability.

Apparently cell probabilities satisfy

$$\sum_{j=1}^J \pi_j = 1, \quad 0 \leq \pi_j \leq 1, \, j \in [J]. \tag{8}$$

The name of cell probability is validated by the following fact.

Theorem 1 *Let* (6) *be CCP distributed with cell probabilities $(\pi_1, \pi_2, \ldots, \pi_J)$. Then*

$$\mathrm{E}(F_j | N = n) = n\pi_j, \quad j \in [J].$$

The most overdispersed case corresponds to the following special case, which is studied by Hoshino (2004, 2005a).

Definition 2 *A CCP distribution* (6) *is called symmetric if all cell probabilities equal $1/J$.*

For example, the CCP distribution generated by $g(z) = z$ amounts to the multinomial distribution. Therefore the family of CCP distributions generalizes the multinomial distribution, while it inherits good properties. Another example is the Dirichlet-multinomial distribution, which is a CCP distribution generated by the logarithmic series distribution. The symmetric Dirichlet-multinomial distribution is used by Takemura (1999) for risk assessment. The CIGP distribution (Hoshino (2003)) is a symmetric CCP distribution generated by an extended (truncated) negative binomial distribution.

On the selection of $g(z)$, Theorem 2 below is helpful. A Modified Power Series (MPS) distribution (Gupta (1974)) over positive integers is defined by the following pgf:

$$g(z) = \sum_{i=1}^{\infty} \frac{a_i \xi^i}{\eta(\xi)} z^i, \tag{9}$$

where

$$\eta(\xi) = \sum_{i=1}^{\infty} a_i \xi^i, \quad 0 \leq a_i,\, 0 < \xi.$$

It should be mentioned that ξ may be some function.

Theorem 2 *Let $F_j, j \in [J]$, be independently compound Poisson distributed as* (4)*, where $g(z)$ is MPS distributed as* (9)*. Then N is sufficient for ξ. In other words, a CCP distribution generated by an MPS distribution* (9) *does not depend on ξ.*

Theorem 2 implies that risk inference based on a CCP distribution is not less exact than that of the unconditional distribution of $\boldsymbol{F}_J$ when F_j is compound Poisson distributed with $g(z)$ of an MPS distribution. Assuming that a risk measure depends on ξ, the error of the estimation of ξ causes inexact risk assessment. Hence it is better that a risk measure does not depend on ξ, which is the case of Theorem 2. This view is formally justified by Rao-Blackwell theorem; see e.g. Lehmann (1991, p.50).

For example, the Poisson-gamma model (Bethlehem et al. (1990)) assumes that F_j is independent and identically compound Poisson distributed with $g(z)$ of the logarithmic series distribution, which is an MPS distribution with $a_i = 1/i$. Bethlehem et al. (1990) let $\mathrm{E}(N)$ equal the observed number of samples; this is equivalent to using a moment estimator, and the uncertainty on ξ affects the risk inference. On the contrary, conditioning on N, the Dirichlet-multinomial distribution eliminates the uncertainty on ξ because of Theorem 2.

Next we see that a CCP distribution is closed under recoding or suppression. To understand this fact, we exploit the reproductivity of a compound Poisson distribution. First we consider unconditionally: $F_j, j \in [J]$, is independently distributed as (4). Then the pgf of a merged frequency $F_1 + \ldots + F_m$ is

$$\exp(\sum_{i=1}^{m} \theta_j (g(z) - 1)),$$

which is again a compound Poisson distribution. More generally, any merged frequency is compound Poisson distributed. In the same course, the pgf of N becomes

$$\exp(\sum_{i=1}^{J} \theta_j (g(z) - 1)).$$

This distribution does not change when cells are merged. As a result, once a CCP distribution is assumed for a contingency table, an anonymized table is also CCP distributed since it is the conditional distribution of independent compound Poisson variables. Similarly, any marginal distribution of a CCP distribution is also a CCP distribution.

Theorem 3 *Let* (6) *be CCP distributed with cell probabilities* $(\pi_1, \pi_2, \ldots, \pi_J)$. *Suppose that* $m \in [J-1]$. *Then for any* m *cells indexed by* $(j_1, j_2, \ldots, j_m)$, *the marginal distribution of* $(F_{j_1}, F_{j_2}, \ldots, F_{j_m}, n - \sum_{l=1}^m F_{j_l})$ *is CCP with cell probabilities* $(\pi_{j_1}, \pi_{j_2}, \ldots, \pi_{j_m}, 1 - \sum_{l=1}^m \pi_{j_l})$.

Theorem 3 entails easy evaluation of the risk of a file or a record. From the definition of a size index (1), we have

$$\mathrm{E}(S_i | N = n) = \sum_{j=1}^J \mathrm{P}(F_j = i | N = n).$$

The right hand side depends on the marginal distribution of F_j, which is bivariate CCP distributed with cell probabilities $(\pi_j, 1 - \pi_j)$ if a CCP distribution is assumed for an entire table. Hence conditioning on N is not very troublesome.

So far, the family of CCP distributions is suitable for modeling a contingency table because (a) it is usable even if covariates are not informative, and (b) it is closed under anonymization techniques. A CCP distribution is more validated when we consider the upper bound of disclosure risk.

The most unsafe case in disseminating microdata is that the most detailed information of individuals is available. This case does not necessarily imply original data without anonymization. For example, an attacker may know more than a statistical agency. If outer databases are matched, an attacker discerns variables that are originally unobserved. This situation is tantamount to the recovery of suppressed variables or the increment of cells of a contingency table. Therefore we would like to know the worst risk for an agency, where J equals infinity.

To investigate this extreme, we consider the sequence of deanonymization: a cell with the highest cell probability is divided since it is regarded as the most coarse part. Repeating this division can be expressed for a CCP distribution by

$$\sum_{j=1}^J \theta_j \to \mu \, (0 < \mu < \infty), \quad \max_j \theta_j \to 0 \quad \text{as } J \to \infty. \tag{10}$$

The first condition of μ in (10) is required to define a cell probability. The limiting distribution of not $\boldsymbol{F}_J$ but $\boldsymbol{S}_n$ is given below because almost every F_j is zero in the limit, where the joint distribution of $\boldsymbol{F}_J$ does not make sense.

Theorem 4 *Suppose that* (6) *is CCP distributed, under which the probability mass function of* $\boldsymbol{S}_n$ *is denoted by* $p_J(\boldsymbol{s}_n)$, *where*

$$\boldsymbol{s}_n \in \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^n i s_i = n, \sum_{i=1}^n s_i \leq J\} =: \mathcal{S}_n(J).$$

Let us define

$$\mathcal{S}_n := \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^n i s_i = n\}.$$

If we apply (10),

$$\lim_{J\to\infty} p_J(\boldsymbol{s}_n) = \frac{n!\mu^u \prod_{i=1}^n {q_i}^{s_i}\frac{1}{s_i!}}{B_n(\mu x_1,\dots,\mu x_n)}, \quad \boldsymbol{s}_n \in \mathcal{S}_n, \tag{11}$$

where $x_i = i!\, q_i, u = \sum_{i=1}^n s_i$*, and* $B_n(\mu x_1,\dots,\mu x_n)$ *is a Bell polynomial defined by*

$$B_n(x_1,\dots,x_n) := n! \sum_{\boldsymbol{s}_n\in\mathcal{S}_n} \prod_{i=1}^n \Big(\frac{x_i}{i!}\Big)^{s_i}\frac{1}{s_i!}. \tag{12}$$

The pmf of size indices $p_J(\boldsymbol{s}_n)$ is obtained by the change of variables (1). In the special case of a symmetric CCP distribution, it becomes

$$p_J(\boldsymbol{s}_n) = \binom{J}{s_0\, s_1\, \cdots\, s_n} \prod_{i=0}^n \mathrm{P}(F_1 = i)^{s_i}\frac{1}{\mathrm{P}(N=n)}, \tag{13}$$

where $s_0 = J - \sum_{i=1}^n s_i$.

We call the right hand side of (11) the Limiting CCP (LCCP) distribution generated by $g(z)$. For example, the LCCP distribution generated by the logarithmic series distribution is the Ewens (1972) distribution. Another LCCP distribution is the Limiting CIGP distribution discussed by Hoshino (2006). The LCCP distribution is a special case of Gibbs partition (Pitman (2006)). For more on a Bell polynomial, see e.g. Charalambides (2002, p.412)

Accordingly, the upper bound of disclosure risk can be evaluated by an LCCP distribution. Since all the cells are homogeneous in the limit, it suffices to evaluate a file level risk measure, which requires the expectation of a size index below.

Theorem 5 *Suppose that* $\boldsymbol{S}_n$ *is LCCP distributed as the right hand side of* (11)*. Then for all* $r_1,\dots,r_n \in \mathbb{N}_0$ *such that* $q := \sum_{i=1}^n ir_i \le n$*, the factorial moments are*

$$\mathrm{E}(\prod_{i=1}^n S_i^{(r_i)}) = \frac{B_{n-q}(\mu x_1,\dots,\mu x_{n-q})\mu^r n^{(q)}}{B_n(\mu x_1,\dots,\mu x_n)} \prod_{i=1}^n (\frac{x_i}{i!})^{r_i}, \tag{14}$$

where $r = \sum_{i=1}^n r_i$ *and* $n^{(q)} = n(n-1)\cdots(n-q+1)$.

3 The quasi-multinomial distribution

This section introduces the QM distribution as a useful CCP distribution. Consul and Mittal (1977) derived the QM distribution by conditioning independent random variables, which is the same as our construction of a CCP distribution.

The Borel distribution (Borel (1942)) has the following pmf:

$$\mathrm{P}(i) = \frac{(\lambda i)^{i-1}}{i!}\exp(-\lambda i), \quad i \in \mathbb{N},\, 0 \le \lambda < 1. \tag{15}$$

This distribution is an MPS distribution (9) with $\xi = \lambda\exp(-\lambda)$. The compound Poisson distribution (4) with $g(z)$ of the Borel distribution is called the Lagrangian Poisson (LP) distribution (Consul and Jain (1973)). Its derivation is rather indirect as seen in Johnson et al. (1993, p.394). See also Consul and Famoye (2006) for the literature of the LP distribution and its relatives.

The LP distribution's pmf is

$$\mathrm{P}(x;\theta,\lambda)=\frac{\theta(\theta+x\lambda)^{x-1}}{x!}\exp(-\theta-x\lambda),\quad x\in\mathbb{N}_0, \tag{16}$$

where $\theta\geq 0$, $0\leq\lambda<1$. This distribution (16) is referred to by $LP(\theta,\lambda)$ in the following. The parameter θ is proportional to the mean. When $\lambda=0$, $LP(\theta,\lambda)$ degenerates into the Poisson distribution with mean θ; λ controls the variance. Negative λ, which produces an improper distribution, is not allowed in the present article.

Now we construct the CCP distribution generated by the Borel distribution. Let $F_j, j\in[J]$, be independently distributed as $LP(\theta_j,\lambda)$. Then the joint conditional distribution (6) is expressed by

$$\binom{n}{f_1 f_2\cdots f_J}\frac{1}{\sum\theta_j(\sum\theta_j+n\lambda)^{n-1}}\prod_{j=1}^{J}\theta_j(\theta_j+f_j\lambda)^{f_j-1}, \tag{17}$$

where $f_j\in\{0,1,\ldots,n\}, \sum f_j=n$. Theorem 2 implies that the QM distribution is independent of $\lambda\exp(-\lambda)$. This independence can be confirmed by reparameterizing as $\theta_j/\lambda=:\tau_j$ for example. If $\lambda=0$, (17) is the multinomial distribution. This fact becomes clear when we reparameterize it by a cell probability (7) and $\beta:=\lambda/\sum\theta_j$ as

$$\binom{n}{f_1 f_2\cdots f_J}\frac{1}{(1+n\beta)^{n-1}}\prod_{j=1}^{J}\pi_j(\pi_j+f_j\beta)^{f_j-1}, \tag{18}$$

where cell probabilities satisfy (8) and β is nonnegative. We observe that the QM distribution (18) is a generalized multinomial distribution with an index β of overdispersion.

When $J=2$ the quasi-multinomial distribution (17) reduces to the quasi-binomial distribution (type 2) proposed by Consul and Mittal (1975). This is the marginal distribution of F_j when $\boldsymbol{F}_J$ is quasi-multinomially distributed. Hence it is used to evaluate record level risk. In particular, Franconi and Polettini type risk (2) is the expectation of $1/F_j$ under the truncated quasi-binomial distribution (type 2). That is,

$$\mathrm{E}(\frac{1}{F_j}|F_j\geq 1,N=n)=\sum_{x=1}^{n}\frac{1}{x}\binom{n}{x}\frac{\pi_j(1-\pi_j)(\pi_j+x\beta)^{x-1}(1-\pi_j+(n-x)\beta)^{n-x-1}}{(1+n\beta)^{n-1}-(1-\pi_j)(1-\pi_j+n\beta)^{n-1}}. \tag{19}$$

This expression (19) seems to require some computation. As in Rinott (2003, p.279), $1/\mathrm{E}(F_j|F_j\geq 1,N=n)$ might approximate to it, though $\mathrm{E}(1/F_j)\geq 1/\mathrm{E}(F_j)$ by Jensen's inequality.

$$\mathrm{E}(F_j|F_j\geq 1,N=n)=\frac{\mathrm{E}(F_j|N=n)}{1-\mathrm{P}(F_j=0)}=\frac{n\pi_j(1+n\beta)^{n-1}}{(1+n\beta)^{n-1}-(1-\pi_j)(1-\pi_j+n\beta)^{n-1}}, \tag{20}$$

because of Theorem 1:

$$\mathrm{E}(F_j|N=n)=n\pi_j,$$

which can be directly derived by Abel's formula; see Charalambides (2002, p.207) for example. Table 1 provides a numerical comparison between (19) and the approximation. Notable under-evaluation is observed when data heavily overdisperse or a cell probability is small. Because these cases are typical in practice, the approximation seems unpromising.

π	β	$\mathrm{E}(1/F \mid F \geq 1, N)$	$1/\mathrm{E}(F \mid F \geq 1, N)$
0.9000	0.0001	0.001111	0.001111
0.8000	0.0001	0.001250	0.001250
0.7000	0.0001	0.001429	0.001429
0.6000	0.0001	0.001668	0.001667
0.5000	0.0001	0.002002	0.002000
0.4000	0.0001	0.002505	0.002500
0.3000	0.0001	0.003343	0.003333
0.2000	0.0001	0.005024	0.005000
0.1000	0.0001	0.010111	0.010000
0.9000	0.0010	0.001112	0.001111
0.8000	0.0010	0.001251	0.001250
0.7000	0.0010	0.001431	0.001429
0.6000	0.0010	0.001671	0.001667
0.5000	0.0010	0.002008	0.002000
0.4000	0.0010	0.002515	0.002500
0.3000	0.0010	0.003365	0.003333
0.2000	0.0010	0.005082	0.005000
0.1000	0.0010	0.010375	0.010000
0.9000	0.0100	0.001126	0.001111
0.8000	0.0100	0.001289	0.001250
0.7000	0.0100	0.001505	0.001429
0.6000	0.0100	0.001806	0.001667
0.5000	0.0100	0.002253	0.002000
0.4000	0.0100	0.002980	0.002500
0.3000	0.0100	0.004351	0.003333
0.2000	0.0100	0.007740	0.005000
0.1000	0.0100	0.024702	0.009999
0.9000	0.1000	0.002825	0.001111
0.8000	0.1000	0.005793	0.001250
0.7000	0.1000	0.010789	0.001428
0.6000	0.1000	0.019455	0.001665
0.5000	0.1000	0.034789	0.001993
0.4000	0.1000	0.061960	0.002472
0.3000	0.1000	0.109001	0.003214
0.2000	0.1000	0.186244	0.004448
0.1000	0.1000	0.302835	0.006654
0.9000	1.0000	0.023490	0.001066
0.8000	1.0000	0.046682	0.001138
0.7000	1.0000	0.070530	0.001216
0.6000	1.0000	0.094983	0.001300
0.5000	1.0000	0.119991	0.001393
0.4000	1.0000	0.145500	0.001494
0.3000	1.0000	0.171459	0.001604
0.2000	1.0000	0.197813	0.001724
0.1000	1.0000	0.224510	0.001855

Table 1: Franconi and Polettini type risk ($N = 1000$)

The marginal variance of F_j is derived by Consul and Mittal (1977) for $\beta > 0$ as

$$\mathrm{V}(F_j|N=n) = n\pi_j \left[\left\{ \frac{(n-1)!}{1+n\beta} \sum_{i=2}^{n} \frac{\pi_j + i\beta}{(n-i)!} \left(\frac{\beta}{1+n\beta} \right)^{i-2} \right\} + 1 - n\pi_j \right].$$

On the upper bound of the risk, the LCCP distribution generated by the Borel distribution is given by Hoshino (2005b, Theorem 1). Its pmf for positive ρ is

$$\mathrm{P}(\boldsymbol{s}_n) = n!\, \rho^{u-1}(\rho+n)^{1-n} \prod_{i=1}^{n} \left(\frac{i^{i-1}}{i!} \right)^{s_i} \frac{1}{s_i!}, \quad \boldsymbol{s}_n \in \mathcal{S}_n, \tag{21}$$

where $\mu = \lambda\rho$ in the limiting argument (10). We call (21) the Limiting Quasi-Multinomial (LQM) distribution. See Hoshino (2005b) for the factorial moments of size indices and more results on this distribution.

The parameter estimation of the asymmetric QM distribution is not discussed in this paper, because it requires a long discussion. Cell probabilities are supposed to be set by regression, and after that, β adjusts overdispersion. The detail of this adjustment is studied in the author's subsequent paper.

4 The symmetric quasi-multinomial distribution

A symmetric CCP distribution is of great concern because it accords with the maximum overdispersion. It is important to examine whether a symmetric CCP distribution can sufficiently describe data or not. Hence this section prepares a few results on the symmetric QM distribution for application.

In the symmetric case of the QM distribution, we can evaluate the joint distribution of sizes indices, using (13). To denote the dependence of $\boldsymbol{S}_n$ on J explicitly, we write $\boldsymbol{S}_{n,J}$ here. Let $\alpha := J\beta$. Then for $0 \leq \alpha$

$$\begin{aligned} &\mathrm{P}(\boldsymbol{S}_{n,J} = \boldsymbol{s}_n | N = n; \alpha) \\ &= \frac{(J-1)!n!}{(J+n\alpha)^{n-1}} \prod_{i=0}^{n} \left(\frac{(1+i\alpha)^{i-1}}{i!} \right)^{s_i} \frac{1}{s_i!}, \quad \boldsymbol{s}_n \in \mathcal{S}_n(J). \end{aligned} \tag{22}$$

We refer to (22) as the symmetric QM distribution.

Theorem 6 *Suppose that size indices are distributed as* (22). *Then for* $r_i \in \mathbb{N}_0, i \in [n]$*, such that* $J \geq \sum_{i=1}^{n} r_i =: r, n \geq \sum_{i=1}^{n} ir_i =: q$,

$$\mathrm{E}(\prod_{i=1}^{n} S_i^{(r_i)} | N = n) = \frac{n!(J-1)!(J-r+(n-q)\alpha)^{n-q-1}}{(J+n\alpha)^{n-1}(n-q)!(J-r-1)!} \prod_{i=1}^{n} \left(\frac{(1+i\alpha)^{i-1}}{i!} \right)^{r_i}.$$

In particular,

$$\mathrm{E}(S_i|N=n) = n^{(i)}(J-1) \frac{(J-1+(n-i)\alpha)^{n-i-1}(1+i\alpha)^{i-1}}{(J+n\alpha)^{n-1} i!}.$$

α	E(S_1)	E(S_2)	E(S_3)	E(S_4)	E(S_5)
0.1	888.03	52.19	2.40	0.10	0.00
1	758.14	94.35	13.90	2.25	0.39
10	288.72	92.00	42.57	23.15	13.78
*100	36.34	13.40	7.38	4.82	3.45
500	7.35	2.71	1.50	0.98	0.71
1000	3.68	1.36	0.75	0.49	0.35
LQM($\rho = 100$)	36.68	13.45	7.40	4.83	3.46

Table 2: $N = 1000, J = 10000$

α	E(S_1)	E(S_2)	E(S_3)	E(S_4)	E(S_5)
0.1	790.35	91.11	8.21	0.64	0.05
1	597.36	126.38	31.67	8.71	2.54
10	160.26	57.66	30.13	18.51	12.44
100	18.22	6.74	3.73	2.44	1.75
500	3.68	1.36	0.75	0.49	0.35
1000	1.84	0.68	0.37	0.25	0.18

Table 3: $N = 1000, J = 5000$

Hence risk measures are easily calculated under the symmetric QM distribution. Table 1,2,3 provide the values of $\mathrm{E}(S_i|N = 1000), i \in [5]$, for $J = 10000, 5000, 2500$. We observe that the expectations moderately depend on J.

In Table 2, the expectations under the LQM distribution (21) with $\rho = 100$ are given for comparison. This case of $\rho = 100$ corresponds to the QM distribution with $\alpha = 100$ because J/α was taken to be ρ in deriving (21). The result suggests that the LQM distribution can substitute for the symmetric QM distribution.

If so, the estimator of ρ is usable as an approximate estimator of J/α. Hoshino (2005b) shows that the ML estimator of the LQM distribution is

$$\hat{\rho} = \frac{U_n - 1}{1 - U_n/n},$$

which leads to an estimator of α:

$$\tilde{\alpha} = \frac{J(n - U_n)}{n(U_n - 1)}. \tag{23}$$

This estimator should be valid when J is large. See Table 7 for an empirical examination.

Now we construct the ML estimation of the symmetric QM distribution. Denoting the log-likelihood of (22) by ℓ, its derivatives are given as

$$\frac{d\ell}{d\alpha} = -(n-1)\frac{n}{J + n\alpha} + \sum_{i=0}^{n} s_i(i-1)\frac{i}{1 + i\alpha},$$

$$\frac{d^2\ell}{d\alpha^2} = (n-1)\left(\frac{n}{J + n\alpha}\right)^2 - \sum_{i=0}^{n} s_i(i-1)\left(\frac{i}{1 + i\alpha}\right)^2.$$

α	$\mathrm{E}(S_1)$	$\mathrm{E}(S_2)$	$\mathrm{E}(S_3)$	$\mathrm{E}(S_4)$	$\mathrm{E}(S_5)$
0.1	630.06	139.85	24.26	3.64	0.50
1	403.60	130.01	49.61	20.79	9.24
10	83.04	31.38	17.23	11.12	7.85
100	9.11	3.37	1.87	1.22	0.88
500	1.84	0.68	0.37	0.25	0.18
1000	0.92	0.34	0.19	0.12	0.09

Table 4: $N = 1000, J = 2500$

Employing these equations, the Newton-Raphson method is applicable. Its starting value can be given by the approximate estimator $\tilde{\alpha}$.

Proposition 2 *The Fisher information* $-\mathrm{E}[d^2\ell/d\alpha^2]$ *is* $O(n)$ *from Theorem 6.*

5 An application

This section fits the symmetric QM distribution to a famous data set from μ-ARGUS software (see e.g. Hundepool (2006)). Once the size of a population is given, we can evaluate disclosure risk. The following argument, however, focuses on the fact that even the symmetric QM distribution reasonably describes the test set, dominating other models in fit.

We use the demo data set of μ-ARGUS, which is actually the set `free1` of the sdcMicro package (Ver. 2.1.0) provided by Templ (2007). We select four key variables and globally recode them, as summarized in Table 5; the column of "Variable" lists the name of key variables, "Categories" lists the number of categories after recoding, and "Recoding" lists the break points of the variable, if any. Then the number of cells J equals 3420, over which $n = 4000$ (sample) individuals are distributed. The maximum frequency of a cell is 67, and there are 855 nonempty cells.

Variable	Categories	Recoding
Region	10	breaks=(0,19,39,59,79,99,119,139,159,190)
Sex	2	–
Age	9	breaks=(1,9,19,29,39,49,59,69,100)
Ageyoung	19	–

Table 5: The detail of anonymization

To the anonymized data, 6 models are fitted: The QM model (QM), the Dirichlet-Multinomial model (DM), the Poisson-Lognormal model (Po-Ln), the LQM model (LQM), the Pitman model (Pitman) and the Ewens model (Ewens). QM is defined by (22), and LQM is (21). The definitions of other models are the same as those in Hoshino (2001). LCCP distributions were proposed to assess the upper bound of disclosure risk, but this section uses instances (LQM and Ewens) just for comparison.

Table 6 shows the fits. The first column lists the names of models, and those fits are evaluated by AIC and $\hat{s_1}$, which is $\mathrm{E}(S_1|N = 4000)$ under the ML estimates of parameters. It is better for $\hat{s_1}$ to be closer to the actual number of sample uniques: $s_1 = 335$.

Table 6 also shows the fits for $J = 10000, 2000$ except for LQM, Pitman and Ewens, which do not depend on J. The reason of varying J is to see the dependence of models on the number of empty cells; s_0 changes, but $\boldsymbol{s}_n$ is fixed. One may claim that true J is smaller than 3420 due to structural zeros. On the other hand, when $J \to \infty$, the limiting distribution of DM is Ewens as that of QM is LQM. Hence the fits of QM and DM should resemble those of LQM and Ewens when $J = 10000$. For each J, the ML estimate of α of QM is tabulated in Table 7. We in Section 4 let J/α be LQM's parameter ρ, whose ML estimate $\hat{\rho}$ is 1086.0. Since $J/\hat{\alpha}$ approaches $\hat{\rho}$, we expect the approximate estimator $\tilde{\alpha}$ of (23) is usable when J is large.

QM fits best, and LQM second in Table 6. It is unlikely that DM or Po-Ln wins against QM for reasonable J in our example. The third is Pitman, and we compare the fits of QM and Pitman in more detail. Table 8 gives the actual sample size indices from s_1 to s_9 along with the fits of Pitman and QM. Figure 1 illustrates these data, where the vertical axis corresponds to $\log(s_i + 1)$ and the horizontal axis corresponds to $i - 1$. The upper tail is truncated at the maximum observed frequency: $i = 67$. The line indicates the actual data, "+" is of Pitman, and "×" is of QM. QM defeats Pitman when i is smaller than around 15. Because smaller cells are vital in risk assessment, this example well demonstrates the importance of the QM distribution.

C language programs used in this section are available together with the size indices data from the author's website:

`http://stat.w3.kanazawa-u.ac.jp/owner/qmulti/qmulti.htm`

	$J = 3420$		$J = 10000$		$J = 2000$	
Model	AIC	$\hat{s_1}$	AIC	$\hat{s_1}$	AIC	$\hat{s_1}$
QM	226.30	346.10	229.24	376.66	239.58	304.05
DM	296.62	275.17	273.84	297.06	336.16	246.53
Po-Ln	881.26	481.82	2422.54	1367.12	489.67	362.73
LQM	234.41	389.13				
Pitman	239.65	365.14				
Ewens	265.42	307.53				

Table 6: The fits of models

J	$\hat{\alpha}$	$J/\hat{\alpha}$
2000	1.3454	1486.56
3420	2.6325	1299.15
10000	8.6729	1153.02

Table 7: The ML estimates of α

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
Actual	335.00	175.00	101.00	58.00	30.00	29.00	13.00	14.00	8.00
Pitman	365.14	135.93	76.94	51.00	36.78	27.94	22.00	17.78	14.66
QM	346.10	146.18	83.02	54.18	38.35	28.62	22.18	17.68	14.39

Table 8: The comparison of the expectations of models

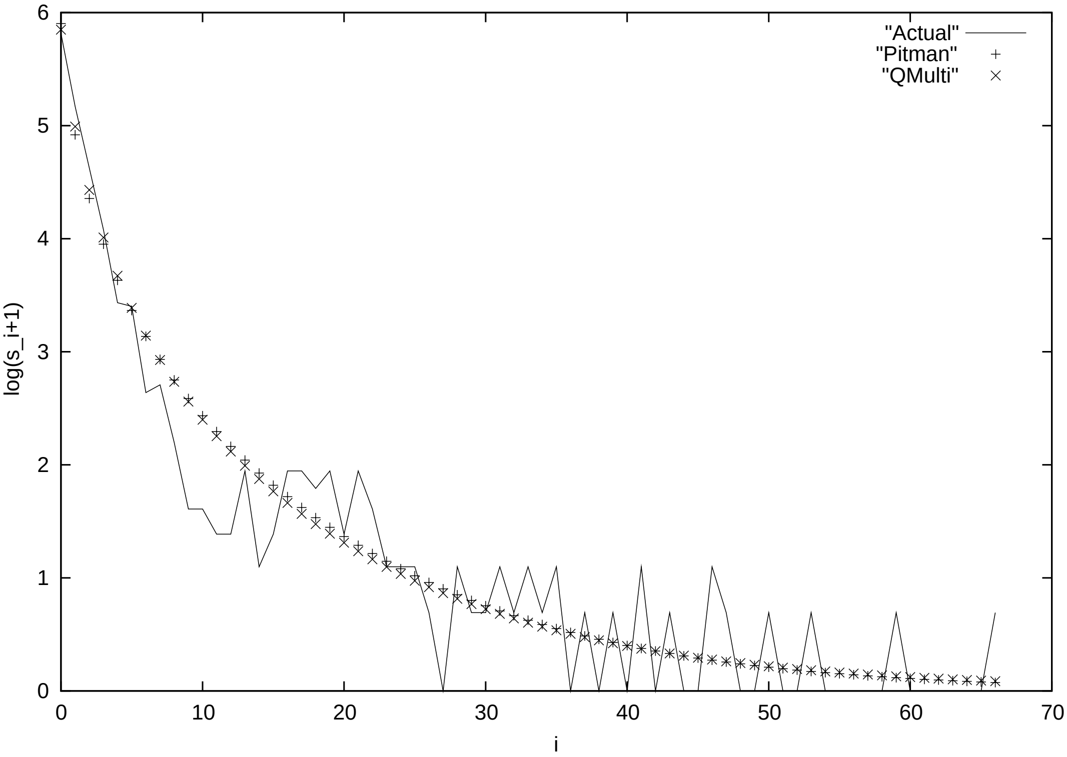

Figure 1: The comparison of the expectations of models

6 Concluding discussion

The negative binomial distribution is basic in modeling overdispersion. Hence, for instance, Franconi and Polettini's risk employs the negative binomial distribution. However, another distribution may be more plausible depending on data; QM was better than DM in Section 5. DM's fit is basically the same as that of the negative binomial distribution, since DM is the conditional distribution of independent negative binomial variables. Therefore considering the family of CCP distributions leads to more precise evaluation of disclosure risk.

Before closing the present article, we think about the safety of local and perturbative anonymization. The CCP distribution was introduced to be consistent with global recoding and suppression, but it is useful for other anonymization in the following sense.

A record of microdata is a vector. Anonymization maps original records $\boldsymbol{x}$'s to masked records $\boldsymbol{y}$'s:

$$\begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} \mapsto \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}.$$

This mapping may or may not be random, but a superpopulation model regards $\boldsymbol{y}_i$ as a random sample. We thus denote the set of possible anonymized expressions by $\mathcal{Y}$ so that $\boldsymbol{y}_i \in \mathcal{Y}$ for all i almost surely. For example, in Section 5, an expression implies the specific combination of Region, Sex, Age and Ageyoung. The number of cells J was 3420, which is $|\mathcal{Y}|$. Generally, the definition of J is not the product of the number of categories but the cardinal of $\mathcal{Y}$. The present paper assumes $\mathcal{Y}$ is countable, which is appropriate. Even if some components of $\boldsymbol{y}$ are continuous variables, they can be treated as discrete since an intruder can not discriminate a subtle difference on the real line.

Many anonymization techniques have been proposed, but there seem to exist only two kinds of safety: (a) Perturbative methods such as noise-addition or swapping regard safety as a distance $d()$ between the original and the anonymized. If $d(\boldsymbol{x}_i, \boldsymbol{y}_i)$ increases, the anonymized set is considered to be safer. (b) Recoding, suppression or microaggregation increases the number of records that have the same expression. If $\sum_{j \neq i} I(\boldsymbol{y}_i = \boldsymbol{y}_j)$ increases, the anonymized set is considered to be safer. Actually, the second type technique has the first safety too.

An existing risk measure, however, may not count both kinds of safety. For example, Mateo-Sanz et al. (2004) enumerate the number of anonymized records that are close to the original using different distances such as standard deviation. This measure neglects the second kind of safety. Another example is the number of population uniques, which does not reflect the first kind. To account for overall safety, Reiter (2005) employs a more complicated approach that evaluates the probability of identification. This method needs estimate the second safety in the process, though.

The CCP distribution is useful for a risk measure that requires population frequencies. The design of anonymization fixes $\mathcal{Y}$, over which individuals are assumed to be CCP distributed. Then the second safety or the population frequency of a given expression $\boldsymbol{y}$ can be estimated, but the deanonymization of $\boldsymbol{y}$ to $\boldsymbol{x}$ is a different problem of the first safety.

In order to understand this fact, let us consider examples. Suppose that a microdata set has two key variables of Sex and Age. Sex has two categories: F and M; Age has 100 categories from 0 to 99. We can add noise to the original records so that $\mathcal{Y} = \{\mathrm{F}, \mathrm{M}\} \times \{0, \ldots, 99\}$. For the i-th observed expression $\boldsymbol{y}_i$, a risk measure may require its population frequency. A CCP distribution can then contribute. As a separate issue, $d(\boldsymbol{x}_i, \boldsymbol{y}_i)$ may be taken into account. If we change the design of anonymization to swapping, $\mathcal{Y}$ can be the same. Then what differs is the structure of $d(\boldsymbol{x}_i, \boldsymbol{y}_i)$, which may or may not affect a risk measure. Consequently, the usefulness of a CCP distribution depends on the selection of a risk measure.

The usefulness of a risk measure, however, depends on the design of anonymization. In particular, local recoding or suppression needs special consideration. Take an example of the two key variables. We first globally recode Age to 3 categories: $\{-14, 15-64, 65-\}$. Then we locally recode an individual $(\mathrm{M}, 15-64)$ to $(\mathrm{M}, 15-)$. As a result, $\mathcal{Y}$ becomes

$$\{(\mathrm{F}, -14), (\mathrm{F}, 15-64), (\mathrm{F}, 65-), (\mathrm{M}, -14), (\mathrm{M}, 15-64), (\mathrm{M}, 65-), (\mathrm{M}, 15-)\} =: \mathcal{Y}_1.$$

Suppose that a population consists of 7 individuals distributed over $\mathcal{Y}_1$ as in Table 9. Two cells of $(\mathrm{M}, 15-64)$ and $(\mathrm{M}, 65-)$ are partially collapsed. We observe that a CCP distribution is still closed under the partial corruption of cells or local recoding (and suppression). Denoting the frequency of $\boldsymbol{y}$ by $F_{\boldsymbol{y}}$,

$$S_1 = \sum_{\boldsymbol{y} \in \mathcal{Y}_1} I(F_{\boldsymbol{y}} = 1) = 3$$

is inappropriate as a risk measure, because the fact that $F_{\boldsymbol{y}} = 1$ does not necessarily imply the uniqueness of an individual. On the other hand, we locally suppress <u>all</u> individuals of $(\mathrm{F}, 65-)$ and $(\mathrm{M}, 65-)$ to $(\mathrm{x}, 65-)$. Then $\mathcal{Y}$ reduces to

$$\{(\mathrm{F}, -14), (\mathrm{F}, 15-64), (\mathrm{x}, 65-), (\mathrm{M}, -14), (\mathrm{M}, 15-64)\} =: \mathcal{Y}_2.$$

The seven individuals are now classified as in Table 10. In this case,

$$S_1 = \sum_{\boldsymbol{y} \in \mathcal{Y}_2} I(F_{\boldsymbol{y}} = 1) = 3$$

is appropriate as a risk measure. The essential difference between $\mathcal{Y}_1$ and $\mathcal{Y}_2$ is that the elements of $\mathcal{Y}$ are mutually exclusive or not. For instance, $(\mathrm{M}, 15-64)$ is an element of $\mathcal{Y}_1$ and included by another element $(\mathrm{M}, 15-)$, but the elements of $\mathcal{Y}_2$ are mutually exclusive.

If only global recoding and suppression are used, the elements of $\mathcal{Y}$ are mutually exclusive. Hence frequency-based risk measures have been proposed for this situation. Actually, other anonymization methods may produce $\mathcal{Y}$ whose elements are mutually exclusive. Then those risk measures are still valid, and a CCP distribution makes sense. To summarize,

Remark 1 *The CCP distribution benefits a frequency-based risk measure, which is inappropriate when possible masked expressions are not mutually exclusive, though.*

Considering the scope of the CCP distribution, we can understand miscellaneous anonymization practices in a unified manner.

<table>
<tr><th>Sex \ Age</th><th>-14</th><th>15-64</th><th>65-</th></tr>
<tr><td>F</td><td>1</td><td>2</td><td>0</td></tr>
<tr><td rowspan="2">M</td><td rowspan="2">2</td><td>0</td><td>1</td></tr>
<tr><td colspan="2">1</td></tr>
</table>

Table 9: Partial local recoding

<table>
<tr><th>Sex \ Age</th><th>-14</th><th>15-64</th><th>65-</th></tr>
<tr><td>F</td><td>1</td><td>2</td><td rowspan="2">1</td></tr>
<tr><td>M</td><td>2</td><td>1</td></tr>
</table>

Table 10: Total local suppression

Acknowledgments

The author would like to express sincere thanks to Prof. Hajime Yamato, Prof. Masaaki Sibuya and Prof. Akimichi Takemura for valuable suggestions on the subject. This research was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Appendix: Proofs

Proof of Theorem 1: First we show the case of $J = 2$ ($N = F_1 + F_2$):

$$\mathrm{E}(F_1|N = n) = n\frac{\theta_1}{\theta_1 + \theta_2}. \tag{24}$$

To show (24), we use this relationship:

$$\frac{dG_1(z)}{dz} = \theta_1 G_1(z)\frac{dg(z)}{dz} = \sum_{x=0}^{\infty}(x+1)\mathrm{P}(F_1 = x+1)z^x. \tag{25}$$

Using the second equation of (25),

$$\theta_1 G_1(z)G_2(z)\frac{dg(z)}{dz} = \sum_{x=0}^{\infty}(x+1)\mathrm{P}(F_1 = x+1)z^x \sum_{y=0}^{\infty}\mathrm{P}(F_2 = y)z^y. \tag{26}$$

The right hand side is rewritten as

$$\sum_{i=0}^{\infty} z^i \sum_{j=0}^{i+1} j\mathrm{P}(F_1 = j)\mathrm{P}(F_2 = i+1-j) = \sum_{i=0}^{\infty} z^i \mathrm{E}(F_1|N = i+1)\mathrm{P}(N = i+1). \tag{27}$$

Then denote the pgf of N by $G(z) = \exp((\theta_1 + \theta_2)(g(z) - 1)) = G_1(z)G_2(z)$. Using the first equation of (25),

$$\frac{\theta_1}{\theta_1 + \theta_2}\frac{dG(z)}{dz} = \theta_1 G(z)\frac{dg(z)}{dz},$$

which is the left hand side of (26). Expanding it further,

$$\frac{\theta_1}{(\theta_1 + \theta_2)}\sum_{x=0}^{\infty}(x+1)\mathrm{P}(N = x+1)z^x$$

has to equal the last expression of (27). By comparing the coefficient of z^{n-1}, (24) is shown.

Now we consider the case of $J \geq 3$. Because of Theorem 3, the marginal distribution of $F_j, j \in [J]$, is a bivariate CCP distribution with cell probabilities $(\pi_j, 1 - \pi_j)$. Therefore the proof of the bivariate case above suffices for the general case, since it does not depend on the index of a cell. Q.E.D.

Proof of Theorem 2: Because $g(z)$ is expressed as $\eta(z\xi)/\eta(\xi)$, we can rewrite $G_j(z)$ as

$$G_j(z) = \exp\left(\frac{\theta_j}{\eta(\xi)}\left(\eta(z\xi) - \eta(\xi)\right)\right) = \frac{\zeta_j(z\xi)}{\zeta_j(\xi)},$$

where $\zeta_j(x) = \exp(\theta_j/\eta(\xi)\eta(x))$. The second equation above implies that F_j is MPS distributed over nonnegative integers. Therefore we can express the joint distribution of frequencies for some $b_{j,i}$ as

$$\mathrm{P}(\boldsymbol{F}_J = \boldsymbol{f}_J) = \xi^n \prod_{j=1}^{J} \frac{b_{j,f_j}}{\zeta_j(\xi)},$$

where $n = f_1 + \cdots + f_J$. Then by the factorization criterion of sufficient statistics (see e.g. Lehmann (1991, p.39)), N is sufficient for ξ. Q.E.D.

Note on Theorem 2 The parameterization of an MPS distribution is not unique. However, parameters are estimated under fixed parameterization. Thus independence from ξ ameliorates risk inference, given specific parameterization.

Proof of Proposition 1: Because $\mathrm{E}(S_i) = J\mathrm{P}(F_j = i)$, (3) equals

$$\frac{\mathrm{P}(F_j = i+1)}{\mathrm{P}(F_j = i)} \geq \frac{\mathrm{P}(F_j = i)}{\mathrm{P}(F_j = i-1)}, \quad i = 1, 2, \ldots.$$

This is a sufficient condition shown by Warde and Katti (1971) for F_j to be a compound Poisson. Q.E.D.

Proof of Theorem 4: This result generalizes Theorem 2.1 of Hoshino (2005a), which concerns a symmetric CCP distribution.

Suppose that $F_j, j \in [J]$, are independently compound Poisson distributed as (4). Then we denote a random vector of size indices by

$$\boldsymbol{T}_{N,J} := (T_1, \ldots, T_N) \stackrel{d}{=} (\sum_{j=1}^{J} I(F_j = 1), \ldots, \sum_{j=1}^{J} I(F_j = N)).$$

The pmf of size indices under a CCP distribution can be expressed as

$$p_J(\boldsymbol{s}_n) = \frac{\mathrm{P}(\boldsymbol{T}_{N,J} = \boldsymbol{s}_n)}{\sum_{\boldsymbol{s}_n \in \mathcal{S}_n(J)} \mathrm{P}(\boldsymbol{T}_{N,J} = \boldsymbol{s}_n)}, \quad \boldsymbol{s}_n \in \mathcal{S}_n(J).$$

We will show under (10) the limit of the numerator of the above equation:

$$\lim_{J\to\infty} \mathrm{P}(\boldsymbol{T}_{N,J} = \boldsymbol{s}_n) = \mu^u \prod_{i=1}^{n} \frac{{q_i}^{s_i}}{s_i!} \exp(-\mu \sum_{i=1}^{n} q_i). \tag{28}$$

If so, the limit of the denominator is

$$\begin{aligned} \lim_{J\to\infty} \sum_{\boldsymbol{s}_n \in \mathcal{S}_n(J)} \mathrm{P}(\boldsymbol{T}_{N,J} = \boldsymbol{s}_n) &= \sum_{\boldsymbol{s}_n \in \mathcal{S}_n} \mu^u \prod_{i=1}^{n} \frac{{q_i}^{s_i}}{s_i!} \exp(-\mu \sum_{i=1}^{n} q_i) \\ &= \frac{\exp(-\mu \sum_{i=1}^{n} q_i)}{n!} B_n(\mu x_1, \ldots, \mu x_n) \qquad (29) \end{aligned}$$

by the definition (12) of a Bell polynomial. Dividing the right hand side of (28) by (29), we have (11). Therefore it suffices to show (28).

Let us denote the pgf of $\boldsymbol{T}_{N,J}$ by

$$G(z_1, \ldots, z_N) = \prod_{j=1}^{J} (1 + \sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i)).$$

Taylor's theorem assures the existence of c_j's satisfying

$$\log G(z_1, \ldots, z_N) = \sum_{j=1}^{J} \sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i) - \frac{1}{2} \sum_{j=1}^{J} \frac{(\sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i))^2}{(1 + c_j)^2},$$

where $0 < c_j < \sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i)$. We bound the last term:

$$\begin{aligned} \sum_{j=1}^{J} \frac{(\sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i))^2}{(1 + c_j)^2} &= \sum_{j=1}^{J} \frac{(\theta_j \sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i)/\theta_j)^2}{(1 + c_j)^2} \\ &\leq \frac{\max_j \theta_j}{(1 + \min_j c_j)^2} \sum_{j=1}^{J} \theta_j (\sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i)/\theta_j)^2. \quad (30) \end{aligned}$$

Hoshino (2005a, eq. B.2) proves that

$$\lim_{\theta_j \to 0} \frac{\mathrm{P}(F_j = i)}{\theta_j} = q_i, \quad i \in \mathbb{N}.$$

Hence by applying (10), the right hand side of (30) goes to zero. Therefore, when we apply (10),

$$\log G(z_1, \ldots, z_N) \to \sum_{j=1}^{J} \theta_j \left\{ \sum_{i=1}^{N} (z_i - 1) \mathrm{P}(F_j = i)/\theta_j \right\} \to \mu \sum_{i=1}^{N} (z_i - 1) q_i.$$

The last expression implies that $T_i, i \in [N]$, is independently Poisson distributed with mean μq_i in the limit. Thus we have shown (28). Q.E.D.

Proof of Theorem 5:

$$\begin{aligned} \mathrm{E}(\prod_{i=1}^{n} S_i{}^{(r_i)}) &= \frac{n! \mu^r}{B_n(\mu x_1, \ldots, \mu x_n)} \sum_{\boldsymbol{s}_n \in \mathcal{S}_n} \prod_{i=1}^{n} \frac{(\mu q_i)^{s_i - r_i} s_i{}^{(r_i)}}{s_i!} (\frac{x_i}{i!})^{r_i} \\ &= \frac{n! \mu^r}{B_n(\mu x_1, \ldots, \mu x_n)} \prod_{i=1}^{n} (\frac{x_i}{i!})^{r_i} \sum_{\boldsymbol{s}_{n-q} \in \mathcal{S}_{n-q}} \prod_{i=1}^{n-q} \frac{(\mu q_i)^{s_i - r_i}}{(s_i - r_i)!} \\ &= \frac{n! \mu^r}{B_n(\mu x_1, \ldots, \mu x_n)} \frac{B_{n-q}(\mu x_1, \ldots, \mu x_{n-q})}{(n-q)!} \prod_{i=1}^{n} (\frac{x_i}{i!})^{r_i}. \end{aligned}$$

Q.E.D.

Proof of Theorem 6:

$$
\begin{aligned}
\mathrm{E}(\prod_{i=1}^{n} S_i^{(r_i)}|N=n) &= \sum_{\boldsymbol{s}_n \in \mathcal{S}_n(J)} \frac{n!(J-1)!}{(J+n\alpha)^{n-1}} \prod_{i=1}^{n} \left(\frac{(1+i\alpha)^{i-1}}{i!} \right)^{s_i} \frac{1}{(s_i-(r_i-1))!} \\
&= \frac{n!(J-1)!}{(J+n\alpha)^{n-1}} \frac{(J-r)(J-r+(n-q)\alpha)^{n-q-1}}{(n-q)!(J-r)!} \prod_{i=1}^{n} \left(\frac{(1+i\alpha)^{i-1}}{i!} \right)^{r_i} \\
&\quad \times \sum_{\boldsymbol{s}_n \in \mathcal{S}_n(J)} \frac{(n-q)!(J-r-1)!}{(J-r+(n-q)\alpha)^{n-q-1}} \\
&\quad \times \prod_{i=1}^{n} \left(\frac{(1+i\alpha)^{i-1}}{i!} \right)^{s_i-r_i} \frac{1}{(s_i-(r_i-1))!}.
\end{aligned}
$$

The summation of the last expression amounts to one because for all $\boldsymbol{s}_{n-q} \in \mathcal{S}_{n-q}(J-r)$, $\mathrm{P}(\boldsymbol{S}_{n-q,J-r} = \boldsymbol{s}_{n-q}|N=n-q)$ is aggregated. Thus we have the result. Q.E.D.

References

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

[2] Borel, E. (1942). Sur l'emploi du théorème de Bernoulli pour faciliter le calcul d'un infinité de coefficients. Application au problème de l'attente à un guichet, *Comptes Rendus, Académie des Sciences, Paris, Series A*, **214**, 452–456.

[3] Charalambides, Ch.A. (2002). *Enumerative Combinatorics*, Chapman and Hall/CRC, New York.

[4] Consul, P.C. and Famoye, F. (2006). *Lagrangian Probability Distributions*. Birkhäuser, Boston.

[5] Consul, P.C. and Jain, G.C. (1973). A generalization of the Poisson distribution. *Technometrics*, **15**, 791–799.

[6] Consul, P.C. and Mittal, S.P. (1975). A new urn model with predetermined strategy. *Biometrische Zeitschrift*, **17**, 67–75.

[7] Consul, P.C. and Mittal, S.P. (1977). Some discrete multinomial probability models with predetermined strategy. *Biometrische Zeitschrift*, **19**, 161–173.

[8] Dale, A. and Elliot, M. (2001). Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. *Journal of the Royal Statistical Society*, A, **164**, 427–447.

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

[10] Forster, J.J. and Webb, E.L. (2007). Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of the Royal Statistical Society*, C, **56**, 551–570.

[11] Franconi, L. and Polettini, S. (2004). Individual risk estimation in μ-ARGUS: a review. In Domingo-Ferrer, J. and Torra, V. (eds). *Privacy in Statistical Databases*, Springer Lecture Notes in Computer Science 3050, Berlin, 262–272.

[12] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

[13] Greenberg, B.V. and Zayatz, L.V. (1992). Strategies for measuring risk in public use microdata file. *Statistica Neerlandica*, **46**, 33–48.

[14] Gupta, R.C. (1974). Modified power series distributions and some of its applications. *Sankhyā*, B, **36**, 288–298.

[15] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, **17**, 499–520.

[16] Hoshino, N. (2003). Random clustering based on the conditional inverse Gaussian-Poisson distribution. *Journal of the Japan Statistical Society*, **33**, 105–117.

[17] Hoshino, N. (2004). Modeling strategy for the risk assessment of privacy. *Proceedings of the 16th RAMP symposium*, 133–148.

[18] Hoshino, N. (2005a). Engen's extended negative binomial model revisited. *Annals of the Institute of Statistical Mathematics*, **57**, 369–387.

[19] Hoshino, N. (2005b). On a limiting quasi-multinomial distribution. *Discussion Paper CIRJE-F-361*, Faculty of Economics, University of Tokyo.

[20] Hoshino, N. (2006). A discrete multivariate distribution resulting from the law of small numbers. *Journal of Applied Probability*, **43**, 852–866.

[21] Hundepool, A. (2006). The ARGUS software in CENEX. In Domingo-Ferrer, J. and Franconi, L. (eds). *Privacy in Statistical Databases*, Springer Lecture Notes in Computer Science 4302, Berlin, 334–346.

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

[23] Lehmann, E.L. (1991). *Theory of point estimation.* Wadsworth, California.

[24] Mandelbrot, B.B. (1983). *The Fractal Geometry of Nature.* W. H. Freeman and Company, New York.

[25] Mateo-Sanz, J.M., Sebé, F. and Domingo-Ferrer, J. (2004). Outlier protection in continuous microdata masking. In Domingo-Ferrer, J. and Torra, V. (eds). *Privacy in Statistical Databases*, Springer Lecture Notes in Computer Science 3050, Berlin, 201–215.

[26] Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness. *Statistical Data Protection - Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 edition*, 59–76, Office for Official Publications of the European Communities, Luxembourg.

[27] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[28] Pitman, J. (2006). *Combinatorial Stochastic Processes*, Lecture Notes in Mathematics 1875, Springer, New York.

[29] Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, **100**, 1103–1112.

[30] Rinott, Y. (2003). On models for statistical disclosure risk estimation. *Monographs of Official Statistics, 3rd Joint ECE/Eurostat Work Session on Statistical Disclosure Control, Luxembourg*, 275–285.

[31] Shlosser, A. (1981). On estimation of the size of the dictionary of a long text on the basis of a sample. *Engineering Cybernetics*, **19**, 97–102.

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

[33] Steutel, F.W. and van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line.* Marcel Dekker, New York.

[34] 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.

[35] Templ, M. (2007). `sdcMicro`: a new flexible R-package for the generation of anonymised microdata – Design issues and new methods. *Joint UNECE/Eurostat work session on statistical data confidentiality, Manchester, United Kingdom, 17-19 December*, WP. 31.

[36] Warde, W.D. and Katti, S.K. (1971). Infinite divisibility of discrete distributions, II. *Annals of Mathematical Statistics*, **42**, 1088–1090.

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

[38] Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics 155, Springer, New York.

[39] Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA.