

ASSESSING THE RISK OF DISCLOSURE BY MODELLING THE NUMBER OF SUB-POPULATIONS OF THE SAME SIZE

Jean-René Boudreau¹ and Patrick St-Cyr²

ABSTRACT

When a statistical agency wants to assess the risk of disclosure of a microdata file, one important measure that has to be estimated is the conditional probability that a record is unique in the population given that it is unique in the sample. The uniqueness of a record is determined by some key variables. The expression of this probability is a function of the sampling fraction and the structure of the population. The structure of a population carries the information on the population in terms of the key variables. The basic problem is to estimate or model the structure of a population. Modelling is preferred when sampling fractions are very small (as is the case in most practical applications). By observing some real populations, we have an idea of the shape of the relationship between the probability and the sampling fraction. In turn, this relationship imposes conditions on the structure. These conditions are important if one wants to fit a model on the structure of a population to assess the risk of disclosure. We will present some of these conditions.

KEY WORDS : Confidentiality; uniqueness; risk of disclosure; microdata.

RÉSUMÉ

Lorsqu'une agence statistique veut évaluer le risque de divulgation dans un fichier de micro-données, elle doit estimer la probabilité conditionnelle qu'un élément soit unique dans la population étant donnée qu'il est unique dans l'échantillon. L'unicité d'une donnée est déterminée au moyen de variables-clé. Cette probabilité dépend de la fraction de sondage et de la structure de la population. La structure de la population comprend l'information sur la population en termes de variables-clé. Le problème fondamental est donc celui de l'estimation ou de la modélisation de la structure de la population. Pour des petites fractions de sondage (le cas le plus courant), la modélisation est préférable. L'expérience avec des populations réelles nous donne une idée du type de relation qui existe entre la probabilité et la fraction de sondage. À son tour, cette relation impose certaines conditions sur la structure. Ces conditions sont importantes lorsqu'il s'agit d'ajuster un modèle à la structure d'une population dans le but d'évaluer le risque de divulgation. Nous présenterons certaines de ces conditions.

MOTS CLÉS: Confidentialité; unicité; risque de divulgation; micro-donnée.

1. INTRODUCTION

Let a population be of size N . This population is partitioned into several sub-populations or classes defined as combinations of values of some discrete variables like age, sex, ethnicity, class of income, etc... These variables will be called key variables. A unique in the population belongs to a sub-population of size one. We select a sample of size n from this population using a simple random sampling plan. A unique in the sample belongs to a class of size one restricted to the

sample. Obviously, a unique in the population that has been selected is also a unique in the sample. The reverse, however, is not always true. We are interested in finding the conditional probability that a unique in the sample is still unique in the population. This evaluation is closely related to the problem of evaluating the risk of disclosure of microdata files released by statistical agencies. In fact, an intruder can match the released file with another file containing direct identifiers like names and addresses using the key variables. If the probability that we are look-

¹ Jean-René Boudreau, Senior Methodologist, Statistics Canada, R.H. Coats, 15-th floor, Ottawa, K1A 0T6.
² Patrick St-Cyr, Methodologist, Statistics Canada, R.H. Coats, 15-th floor, Ottawa, K1A 0T6.

ing for is high, the intruder can be confident that the one to one matches he or she gets, are good ones. The formulation of this probability depends on the sampling fraction, $f = n/N$, and on components of a special vector called the structure of the population. Its i -th component is the number of sub-populations of size i ($i = 1, \dots, N$). The structure of the population will be denoted by $\mathbf{U} = (U_1, \dots, U_N)^T$. For example, we have $N = U_1 + 2U_2 + \dots + NU_N$.

Since we observe only the sample, we have to estimate or model the components of \mathbf{U} . The finite population theory only gives unreliable estimates. One way out is to assume that the population is a realisation of a super-population model. Several authors have tried different models. For example, Bethlehem and al., (1990) used a Poisson-Gamma model. Skinner and Holmes (1992) worked with a Poisson-Lognormal model. The relationship between the probability and the sampling fraction for real populations implies constraints over \mathbf{U} . These constraints might give us some clues to what models should be considered. This paper takes this approach.

2. FORMULATION OF THE CONDITIONAL PROBABILITY

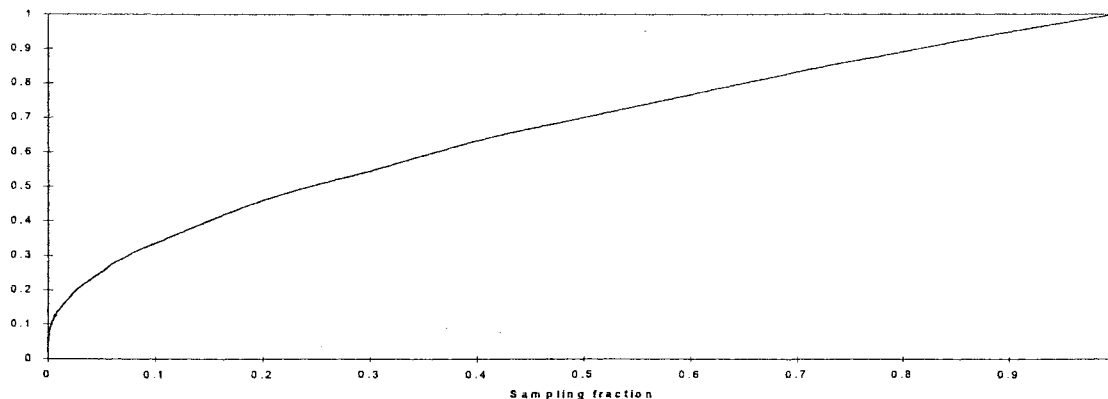
We have a population with N units. We define the « content » of that population as its frequency table based on the key variables. The content parti-

tions the population into s , say, sub-populations of size N_1, \dots, N_s . The « structure of the population » is the vector $\mathbf{U} = (U_1, \dots, U_N)^T$ where $U_j = \text{card} \{ k : N_k = j \}$. We take a sample of size n from that population using a simple random sampling plan. We observe the random vector $\mathbf{n} = (n_1, \dots, n_s)^T$ where n_i is the number of units in the sample from population i . The « structure of the sample » is the random vector $\mathbf{u} = (u_1, \dots, u_n)^T$ where $u_j = \text{card} \{ k : n_k = j \}$. U_1 is the number of uniques in the population and u_1 is the number of uniques in the sample.

Let $u_1 > 0$. Given that an element is unique in the sample, the probability that it is still unique in the population is the ratio $P = U_1/N : u_1/n = f(U_1/u_1)$, where f is the sampling fraction n/N . So we must find an expression of U_1 in terms of the structure of the sample. Figure 1 shows the relationship between P and f for a real population ($N = 781,825$ and five key variables). When $f = 1/N$, then $u_1 = 1$, so $P = U_1/N$ which is the proportion of uniques in the population. When $f = 1$, then obviously $P = 1$. Two things are worth noting. The probability is always higher than the sampling fraction and the relationship is concave. This pattern is always present for real populations.

Figure 1

Conditional probability



Can we estimate this probability using only the sample? The first result in this direction is

$$E\{u_j\} = \sum_{i=0}^{N-n} \frac{C_j^{i+j} C_{n-j}^{N-j-i}}{C_n^N} U_{i+j}.$$

The expectation is calculated according to the sampling plan and the C_i^j 's are the binomial coefficients. The proofs of this formula and the following ones can be found in Boudreau (1995). This equation ties up both structures, in particular, the two kinds of uniqueness. For $j = 1$, we have

$$E\{u_1\} = \sum_{i=1}^{N-n+1} \frac{C_{n-1}^{N-i}}{C_n^N} i U_i.$$

From these equations, if we suppose that the sample size is always as large as any sub-population sizes (which is true if the content is rich), we get an unbiased estimate of the components of U . In fact, we find that

$$\hat{U}_j = \frac{C_n^N}{C_{n-j}^{N-j}} \sum_{i=0}^{n-j} (-1)^i \frac{C_j^{i+j} C_i^{N-n+i-1}}{C_i^{n-j}} u_{i+j}$$

is an unbiased estimate of U_j . In particular, for $j = 1$, we have

$$\hat{U}_1 = \frac{N}{n} \sum_{i=1}^n (-1)^{i-1} \frac{C_{i-1}^{N-n+i-2}}{C_{i-1}^{N-1}} i u_i.$$

Therefore if $u_1 > 0$, an estimator of P can be formulated as

$$\hat{P} = \sum_{i=1}^n (-1)^{i-1} \frac{C_{i-1}^{N-n+i-2}}{C_{i-1}^{n-1}} \frac{i u_i}{u_1}.$$

However, this estimator is very unstable or chaotic for small n . This is understandable: it is not possible for a sample to carry all the information of the structure of a population. Because of this, we are urged to model the structure of a population.

3. CONDITIONS ON THE STRUCTURE OF THE POPULATION

We want to generate structures of a population using a model. It may be preferable to consider, instead of P (the conditional probability), the following quantity

$$f \frac{U_1}{E\{u_1\}}.$$

This quantity does not depend on particular sample realisations. As usual, the expectation is computed according to the sampling plan. We will still say, for the sake of this exercise, that this quantity is the conditional probability under consideration and will be denoted by the same letter P . In this section, we will try to find some conditions on the structure of a population that would give similar behaviours for P .

If we rearrange the expression of the previous expectation in terms of the structure of the population, we get

$$E\{u_1\} = \sum_{i=1}^{N-n+1} \frac{n}{N-i+1} \left(1 - \frac{n}{N}\right) \cdots \left(1 - \frac{n}{N-i+2}\right) i U_i.$$

So, P is then approximated by

$$P \approx \frac{1}{\sum_{i=1}^{N(1-f)+1} (1-f)^{i-1} \frac{i U_i}{U_1}}.$$

This approximation is very good if N is large and f small, which is true in most practical applications. Define the « probability distribution generated by a structure of the population » by the vector $\mathbf{p} = (p_1, p_2, \dots, p_N)^T$ where $p_i = i U_i / N$. That is, p_i is the probability to belong to a sub-population of size i ($i = 1, \dots, N$). Then P can be rewritten as

$$P = p_1 \frac{1-f}{E\{(1-f)^X\}}.$$

Here, the random variable X follows the distribution \mathbf{p} . Let $\phi_r = \phi_r(f) = E\{X^r (1-f)^X\}$ for $r = 0, 1, \dots$ With this notation, we have $P = p_1 (1-f) / \phi_0$. The Taylor development of P around f_0 is

$$P = \frac{p_1}{\phi_0} \left[1 - f_0 + \frac{\phi_1 - \phi_0}{\phi_0} (f - f_0) + \frac{2\phi_2^2 - \phi_1\phi_0 - \phi_2\phi_0}{2(1-f_0)\phi_0^2} (f - f_0)^2 \right] + o(|f - f_0|^3).$$

The ϕ 's are evaluated at f_0 . In particular, for $f_0 = 0$, we have

$$P = p_1 \left[1 + (\mu - 1)f + \mu^2(1 - 1/\mu - vc^2)f^2 \right] + o(|f|^3).$$

μ and cv are the mean and the coefficient of variation of the distribution \mathbf{p} . We have seen that the conditional probability is a concave function of the sampling fraction. That means in particular that, for $f_0 = 0$, we have $cv^2 \geq 1 - 1/\mu$. That is, the distribution \mathbf{p} should have a large variance. There must therefore exist some sub-populations with large sizes. The concavity of the relationship also implies that $2\phi_1^2 - \phi_1\phi_0 - \phi_2\phi_0 < 0$. This condition also restricts the choices for \mathbf{p} .

What is the shape of all points \mathbf{p} that satisfy the last inequality? To answer this question, we will use finite differences to find an equivalent relation in the finite case.

3.1 Quadratic relations related to \mathbf{P}

We will take the exact formulation instead of the approximation. Therefore, we will use finite differences instead of derivatives. We write $P_n = U_1/Q_n$ where

$$Q_n = \sum_{i=1}^{N-n+1} \frac{C_{n-1}^{N-i}}{C_{n-1}^{N-1}} i U_i,$$

$$\begin{aligned} \Theta &= \sum_{i=1}^{N-n} \sum_{j=1}^{N-n} \left[\alpha_{i,n-1} (\alpha_{j,n} - \alpha_{j,n+1}) - \alpha_{i,n+1} (\alpha_{j,n-1} - \alpha_{j,n}) \right] F_i F_j \\ &\quad + \sum_{i=N-n+1}^{N-n+2} \sum_{j=1}^{N-n} \left[\alpha_{i,n-1} (\alpha_{j,n} - \alpha_{j,n+1}) \right] F_i F_j \\ &\quad + \sum_{i=1}^{N-n} \left[\alpha_{i,n-1} \alpha_{N-n+1,n} - \alpha_{i,n+1} (\alpha_{N-n+1,n-1} - \alpha_{N-n+1,n}) \right] F_i F_{N-n+1} \\ &\quad - \sum_{i=1}^{N-n} \left[\alpha_{i,n+1} \alpha_{N-n+2,n-1} \right] F_i F_{N-n+2} + \alpha_{N-n+1,n-1} \alpha_{N-n+1,n} F_{N-n+1}^2 + \alpha_{N-n+2,n-1} \alpha_{N-n+1,n} F_{N-n+2} F_{N-n+1}. \end{aligned}$$

After noticing that

- $\alpha_{i,n-1} = \frac{N-n+1}{N-n-i+2} \alpha_{i,n}$ ($i = 1, \dots, N-n+1$),
- $\alpha_{i,n+1} = \frac{N-n-i+1}{N-n} \alpha_{i,n}$ ($i = 1, \dots, N-n+1$),

($n = 1, \dots, N$). The first and the second finite differences at the point n are respectively $\Delta_n = P_{n+1} - P_n$ for $n = 1, \dots, N-1$ and $\Delta_n^2 = P_{n+2} - 2P_{n+1} + P_n$ for $n = 1, \dots, N-2$. We know that P is concave at the point n if, and only if, $\Delta_{n-1}^2 < 0$. This inequality is equivalent to $\Theta = \Theta_n = Q_{n-1}(Q_n - Q_{n+1}) - Q_{n+1}(Q_{n-1} - Q_n) < 0$. The next step is to express this relation in terms of a matrix operator.

First, we split the components of Q_n . Let $\alpha_{i,n}$ be the ratio of the two binomial coefficients and \bar{F}_i be the remaining terms. We then have

$$Q_n = \sum_{i=1}^{N-n+1} \alpha_{i,n} F_i.$$

Our aim is to express Θ as $\Theta = \mathbf{F}^T \mathbf{A} \mathbf{F}$, where $\mathbf{F} = (F_1, \dots, F_{N-n+2})^T$ and \mathbf{A} (which depends on n) is a square matrix of dimension $N-n+2$. If we calculate Θ , we obtain

we can isolate $\alpha_{i,n} \alpha_{j,n}$ for all entries $i, j \neq N - n + 1$. After symmetrisation, the matrix has the following form

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{C}^T & -\frac{1/2}{C_{n-2}^{N-1}} \\ \mathbf{C} & \mathbf{B} & \mathbf{D} \\ -\frac{1/2}{C_{n-2}^{N-1}} & \mathbf{D}^T & 0 \end{bmatrix},$$

where

- $\mathbf{B} = \left[\frac{\alpha_{i,n} \alpha_{j,n}}{(N-n-i+2)(N-n-j+2)} \left\{ \left(1 + \frac{1}{N-n} \right) \left(\frac{-(i-j)^2}{2} + (i-1)(j-1) + \frac{i+j}{2} - 1 \right) - \frac{1}{N-n} (i-1)(j-1) \frac{i+j}{2} \right\} \right]_{\substack{i=2, \dots, N-n+1 \\ j=2, \dots, N-n+1}}$
- $\mathbf{D} = \left[-\frac{\alpha_{j,n}}{2C_{n-2}^{N-1}} \left\{ 1 - 2 \frac{j-1}{N-n} \right\} \right]_{j=2, \dots, N-n+1}^T$
- $\mathbf{C} = \left[-\frac{\alpha_{j,n}}{2} \frac{j-1}{N-n} \frac{j-2}{N-n-j+2} \right]_{j=2, \dots, N-n+1}^T$

These matrices depend only on the sampling plan. If we divide all terms by N^2 , we obtain the following formulation for the concavity condition

$$\Theta = \mathbf{p}_{\diamond}^T \mathbf{B} \mathbf{p}_{\diamond} + 2(p_1 \mathbf{C}^T + p_{N-n+2} \mathbf{D}^T) \mathbf{p}_{\diamond} - \frac{p_1 p_{N-n+2}}{C_{n-2}^{N-1}} < 0 \quad (1)$$

Here $\mathbf{q} = \mathbf{q}_n = (p_1, \mathbf{p}_{\diamond}, p_{N-n+2})$. We can easily verify that \mathbf{A} is not negative definite (\mathbf{A} is in fact indefinite), so $\Theta < 0$ does not hold for all \mathbf{q} .

Since the vector \mathbf{q} has non-negative components, a search for non-positive matrix elements and their magnitude is necessary. The most significant matrix in the condition is \mathbf{B} . It has some negative terms. They are situated at the top right and at the bottom left corners. This is due to the term $(i-j)^2$ in the definition of the elements of \mathbf{B} . It means that this matrix will generate negative values when the terms in the upper right corner are large, and when \mathbf{p}_{\diamond} has non-negligible components situated at the extremities. Now, we look at the translation part of the quadratic

form. First, all elements of \mathbf{C} are non-positive. Also, only elements d_j where j is greater than $(N-n)/2$ are positive in the matrix \mathbf{D} . In real situations (that is, when f is small, U_1 is dominant and U_{N-n+2} is small), $p_1 \mathbf{C}^T + p_{N-n+2} \mathbf{D}^T$ should have all negative elements. The translation part will contribute negative terms to the sum.

In summary, we find that P is concave at the point n/N if:

- the sampling plan specifies large elements in the upper right corner of the matrix \mathbf{B} ,
- the vector \mathbf{q} has non-negligible components situated near the ends;
- p_1 is dominant.

Table 1 gives three portions of the structure of the population used to draw the previous figure.

Table 1

i	U _i	p _i	i	U _i	p _i	i	U _i	p _i
1	35718	0.0457	196	6	0.0015	3926	1	0.0050
2	9549	0.0244	197	4	0.0010	4016	1	0.0051
3	4421	0.0170	198	3	0.0008	4082	1	0.0052
4	2642	0.0135	199	2	0.0005	4280	1	0.0055
5	1745	0.0111	200	1	0.0002	4407	1	0.0056
6	1321	0.0101	201	4	0.0010	4563	1	0.0058
7	983	0.0088	202	4	0.0010	4619	1	0.0059
8	805	0.0082	203	3	0.0005	5855	1	0.0075
9	643	0.0073	204	2	0.0007	6447	1	0.0082
10	537	0.0069	205	7	0.0018	6604	1	0.0084
11	453	0.0064	206	3	0.0007	9881	1	0.0126
12	377	0.0058	207	1	0.0003	9941	1	0.0127
13	360	0.0060	208	3	0.0008	13705	1	0.0175
14	306	0.0055	209	2	0.0005	28491	1	0.0364

\mathbf{p} shows exactly the kind of probability distributions required to have a concave relation between P and f .

The relation (1) only gives a very broad description of \mathbf{q} . A deeper analysis of the shape requires an expression of (1) in terms of elementary relations. That will be done in the next sub-section.

3.2 Fundamental shape of \mathbf{q}

We want to rewrite the quadratic form $\mathbf{q}^T \mathbf{A} \mathbf{q}$ as a sum of squares. Every symmetric matrix has its own spectral decomposition. Therefore \mathbf{A} can be decomposed into the following sum of squares

$$\mathbf{A} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + \dots + \lambda_{N-n+2} \mathbf{u}_{N-n+2} \mathbf{u}_{N-n+2}^T \quad (2)$$

where λ_i and \mathbf{u}_i are respectively the i -th eigenvalue and the corresponding eigenvector of \mathbf{A} . Table 2 gives eigenvalues for different N and n . Although the values of N and n shown in the table are not realistic,

they illustrate the variations of the eigenvalues. We observe that only the first, second and last eigenvalues are significantly different from 0. The absolute value of others eigenvalues are not greater than 10^{-16} .

Table 2

	N=50		N=100		N=200	
	n=5	n=25	n=5	n=25	n=5	n=25
λ_1	.162	.0020	.302	.0025	.583	.0041
λ_2	.081	.0010	.150	.0013	.288	.0021
$\lambda_3, \dots, \lambda_*$	$+o(10^{-17})$	$+o(10^{-19})$	$+o(10^{-17})$	$+o(10^{-18})$	$+o(10^{-17})$	$+o(10^{-18})$
$\lambda_*, \dots, \lambda_{N-n+1}$	$-o(10^{-17})$	$-o(10^{-19})$	$-o(10^{-17})$	$-o(10^{-19})$	$-o(10^{-17})$	$-o(10^{-18})$
λ_{N-n+2}	-.082	-.0010	-.153	-.0013	-.295	-.0021

* is the largest integer less than or equal to the value $(N-n+2)/2$

Denote \mathbf{u}_1 , \mathbf{u}_2 and \mathbf{u}_{N-n+2} by \mathbf{u} , \mathbf{v} and \mathbf{w} respectively. The results in Table 2 motivates the following approximation:

$$\mathbf{A} \approx \lambda_1 \mathbf{u}\mathbf{u}^T + \lambda_2 \mathbf{v}\mathbf{v}^T - |\lambda_{N-n+2}| \mathbf{w}\mathbf{w}^T.$$

We now get a new formulation for the concavity condition:

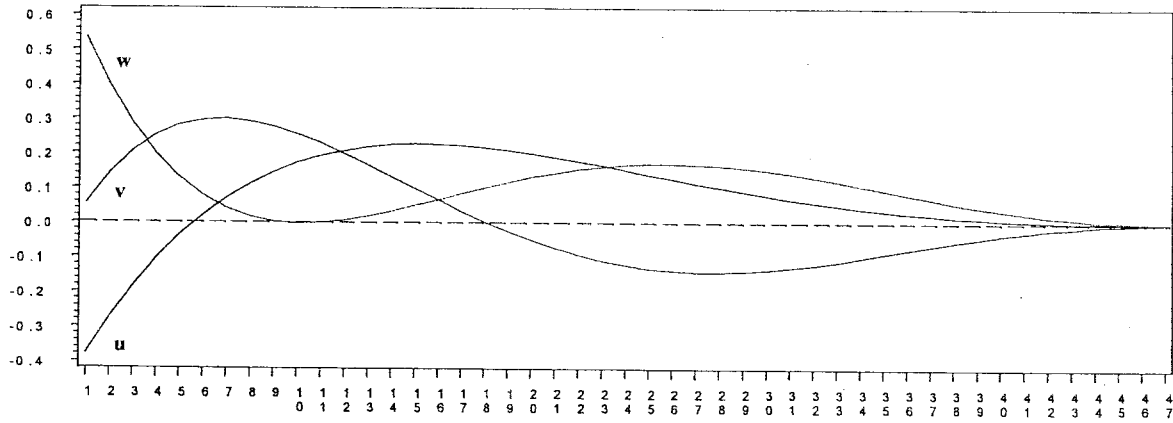
$$\lambda_1 (\mathbf{q}^T \mathbf{u})^2 + \lambda_2 (\mathbf{q}^T \mathbf{v})^2 - |\lambda_{N-n+2}| (\mathbf{q}^T \mathbf{w})^2 < 0 \quad (3)$$

The condition (3) will be closely satisfied for small values of $|\mathbf{q}^T \mathbf{u}|$ and $|\mathbf{q}^T \mathbf{v}|$ and high values of $|\mathbf{q}^T \mathbf{w}|$. Since $\mathbf{u}^T \mathbf{w} = \mathbf{v}^T \mathbf{w} = 0$, the minimum of the

left member of (3) is achieved when $\mathbf{q} = \mathbf{v}\mathbf{w}$ where v is a positive constant to ensure that \mathbf{q} belongs to the simplex $\mathbf{1}^T \mathbf{q} = 1 - \rho$ with $\rho = p_{N-n+3} + \dots + p_N$. We observed, from many examples, that $\mathbf{w} \geq \mathbf{0}$ when the sampling fraction is small. Since $\mathbf{w}^T \mathbf{w} = 1$, the minimum is $-v^2 \lambda_{N-n+2}$.

Figure 2 gives the distributions \mathbf{u} , \mathbf{v} and \mathbf{w} for the case $N=50$ and $n=5$. The distribution \mathbf{w} gives us the kind of shape for \mathbf{q} that we expected. That is, significant components situated at the left side with p_1 dominant and decreasing rapidly followed by a long tail. The distributions \mathbf{u} and \mathbf{v} look respectively like the opposite of the left side and the right side of \mathbf{w} . Small contributions coming from \mathbf{u} or \mathbf{v} will reduce either p_1 or the tail.

Figure 2



We want to describe the set of solutions \mathbf{q} in terms of \mathbf{u} , \mathbf{v} and \mathbf{w} . The key is to understand the geometric meaning of (3) when \mathbf{q} lies in the simplex. First, \mathbf{q} is related to a conic. Suppose $\mathbf{q} = \mathbf{x}\mathbf{u} + \mathbf{y}\mathbf{v} + \mathbf{z}\mathbf{w}$. Then (3) implies $\lambda_1 x^2 + \lambda_2 y^2 - |\lambda_{N-n+2}| z^2 < 0$, and its graph is the interior of an elliptic cone centred at the origin. The major axis of each ellipse at level surface z (for any choice of $z > 0$) has half-length $\gamma_{N-n+2} \gamma_1^{-1} z$, and the minor axis has half-length $\gamma_{N-n+2} \gamma_2^{-1} z$ with $\gamma_i = |\lambda_i|^{1/2}$. So that: $|x| < \gamma_{N-n+2} \gamma_1^{-1} z$ and $|y| < \gamma_{N-n+2} \gamma_2^{-1} z$. Furthermore, these axes point in the direction of \mathbf{u} and \mathbf{v} respectively. As

$$\lambda_1 x^2 + \lambda_2 y^2 - |\lambda_{N-n+2}| z^2 = 0. \quad (4)$$

The parametric equations of (4) are: $x = \gamma_{N-n+2} \gamma_1^{-1} z \cos\theta$ and $y = \gamma_{N-n+2} \gamma_2^{-1} z \sin\theta$, with θ varying between 0 and 2π .

the contribution of \mathbf{w} gets larger, the resulting \mathbf{q} can deviate more from \mathbf{w} in the direction of \mathbf{u} and \mathbf{v} . The set of solutions is in the intersection of the elliptic cone and the simplex related to \mathbf{q} . This intersection forms an ellipse truncated by the sides of the simplex. The simplex implies that the domain of z is now restricted to an interval.

The algebraic expression of the truncated ellipse is not easy to determine. That is why we restrict our attention to a specific example ($N=50$, $n=5$). To illustrate the set of solutions for the case $N=50$ and $n=5$, we will refer to the limiting case

The method used to generate admissible \mathbf{q} is simple: we looped on (z, θ) and selected $\mathbf{q} = \mathbf{x}\mathbf{u} + \mathbf{y}\mathbf{v} + \mathbf{z}\mathbf{w}$ which are on the simplex and satisfy (3). The most representative distributions are given in Figure 3

(without loss of generality, we assume $\rho = 0$). The set of solutions lies inside the range defined by these distributions.

As shown in Figure 3, distributions (a) to (e) vary according to p_1 and the tail of the distribution.

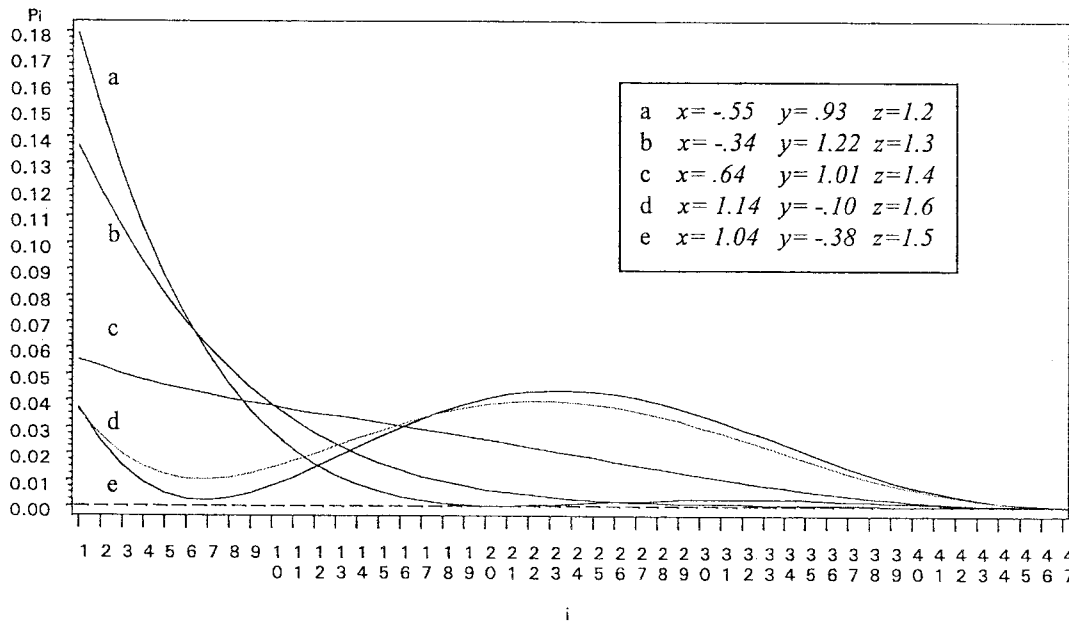
- In (a), a very dominant left side components with a negligible long tail,

in (b), the tail appears and the left side component values decrease,

- in (c), the distribution decreases slowly with an inflection point at p_7 ,
- in (d) and (e), p_1 and the tail are closed to their minimal and maximal value respectively.

In all cases, we observe that the distribution seems to be a mixture of two fundamental distributions.

Figure 3



4. MODELLING THE STRUCTURES OF POPULATIONS

From the results of the last section, we will be successful in generating structures of a population if the model can create lots of uniques in the population and some large sub-populations. We study one of the models considered in the literature : the Poisson-Gamma model.

Let a population of size N be partitioned into s different sub-populations. Let Z_i be the random vari-

able describing the number of elements of the sub-population i ($i = 1, \dots, s$). Assume that all Z_i are mutually independent and they follow a Poisson distribution with parameter $\mu_i = N\pi_i$ (π_i is the probability to belong to the sub-population i). Assume moreover that π_i ($i = 1, \dots, s$) are realisations of independent and identically random variables of the gamma type with parameters α and β . Then, from the model, $E[\pi] = \alpha\beta$ and, since $\sum_i E[\pi] = 1$, we have $s\alpha\beta = 1$. With all these assumptions, we have

$$P(Z_i = 1) = N\alpha\beta (1 + N\beta)^{-(1+\alpha)} = \frac{N}{s} (1 + N\beta)^{-(1+\alpha)},$$

and the expected proportion of uniques in the population is

$$p_1 = \frac{U_1}{N} = \frac{sP(Z_i = 1)}{N} = (1 + N\beta)^{-(1+\alpha)}.$$

The conditional probability P equals, if we denote $(\beta N)^{-1}$ by γ , the following expression

$$P = f \frac{U_1}{u_1} = \left(\frac{f + \gamma}{1 + \gamma} \right)^{1+\alpha}.$$

The relation between P and f is, however, convex because the condition $\alpha > 0$ is required by the model. The model does not generate enough large sub-populations. The Poisson-Gamma model is not that far from a satisfactory solution. If we try to directly model the relationship between P and f , the expression

$$P = \left(\frac{f + \gamma}{1 + \gamma} \right)^\alpha$$

where $0 < \alpha < 1$ and $\gamma > 0$ give a very good fit. This relationship is concave. If we start from this relation and assume that P is the ratio of the proportions of the two kinds of uniqueness, we find that $p_1 = (1 + N\beta)^{-\alpha}$ where $0 < \alpha < 1$ and $\beta > 0$. We presume that this model fits the left side of \mathbf{p} adequately but it is unable to generate large tails. If we consider, however, mixtures like $Z = \varepsilon X + (1-\varepsilon)Y$, where

- a) X is a random variable governed by a Poisson-Gamma distribution, X would generate lots of small sub-populations;
- b) Y is a random variable that follows a distribution that would generate large sub-populations, the normal or Cauchy distribution would be a good candidate as seen in figure 2;
- c) ε is a real parameter between 0 and 1,

then we may be in a much better position when we estimate these parameters from the sample. We might get enough degrees of freedom to fit these kinds of models adequately to the real world.

5. CONCLUSION

In this paper, we have studied the conditional probability of being unique in the population given the fact of being unique in the sample. This statistic is central to the evaluation of the risk of disclosure of microdata files. By observing the relationship between this probability and the sampling fraction for real populations, we were able to find properties that should be useful at the modelling stage. The strong result in the paper conjectures that a mixture of two distributions is the key to adequately fit structures of real populations.

REFERENCES

- Bethlehem, J. G, Keller, W.J., Pannekoek, J. (1990). *Disclosure Control of Microdata*, JASA, 85, pp. 38-45.
- Boudreau J. R. (1995). Assessment and Reduction of Disclosure Risk in Microdata Files Containing Discrete Data. *Proceedings of Statistics Canada Symposium 95*. Canada. pp. 143 – 153.
- Skinner C. J. Holmes D. J. (1992). Modelling Population Uniqueness. *International Seminar on Statistical Confidentiality*. Dublin