Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 373–383

A Bayesian, Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment

Stephen M. Samuels[1]

One important measure of disclosure risk for microdata is the proportion of sample uniques which are also population uniques. The distribution of this random variable depends on the population only through its partition structure: the distribution of the numbers of cells of each size. Partition distributions have been extensively studied in population genetics. Portions of that research can be adapted to provide us with the promise of a mathematical framework based on plausible prior distributions with easy to interpret parameters, and a modified Polya urn sampling model from which risk assessment is easily obtained.

Key words: Partition structure; Polya urn; Poisson-Dirichlet.

1. Introduction

Recognizing that zero-risk requirements for disclosure of statistical records are, in practice, impossibly high standards, the Panel on Confidentiality and Accessibility in Government Statistics recommended, in Duncan, Jabine, and de Wolf (1993), release of information for legitimate statistical purposes that entail a reasonably low risk of disclosure of individually identifiable data. This naturally raises the question of how to measure disclosure risk. The first item in the Research Agenda of U.S. Office of Management and Budget's (OMB) Statistical Policy Working Paper 22 (Kirkendall et al. 1994) is the following:

> The definition and the assessment of disclosure risk in microdata need to be put on a sound statistical footing. Probability theory provides an intuitively appealing framework for defining disclosure in microdata in which we relate disclosure to the probability of reidentification. Without a measure of disclosure risk, decisions concerning the disclosure limitation of microdata files must be based on precedents and judgment calls. Research into probability-based definitions of disclosure in microdata should have high priority.

The report continues: ''One part of this research involves developing a method of estimating the per cent of records on a sample microdata file that represent unique persons or establishments in the population.'' This is a version of what is sometimes called *the uniques problem*.

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907-1399, U.S.A. A somewhat earlier version was presented at SDP'98 in Lisbon, with the title, ''A Bayesian, Population-Genetics-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment.'' Support for this research from the U.S. Bureau of the Census is gratefully acknowledged.

The canonical form of the uniques problem (Chen and Keller-McNulty 1996, de Waal and Willenborg 1996, Fienberg and Makov 1996, Skinner et al. 1994 and others) is this: there is a microdata file containing, say, N records, each of which refers to a person or establishment, and consists of a list of the values, for that person or establishment, for each of several categorical variables. When the records are cross-classified according to a subset of these variables, the database is thereby partitioned into, say, K non-empty cells, a_j of which are of size j., for $j = 1, 2, \ldots$. Those of size one are the *population uniques*. A random sample of size n is to be selected, and, presumably, made public. Any population uniques which are included in the sample – where they are, of course, *sample uniques* – may constitute an unacceptable disclosure risk, unless they are swamped by relatively large numbers of sample uniques which are *not* population uniques.

One important measure of disclosure risk for such microdata is the random variable, p: the proportion of sample uniques which are also population uniques. If that proportion is low, then an intruder, attempting to make an identification on the basis of a sample unique, is very likely to be making a *false* identification. (This, in itself, is a kind of disclosure risk that must be addressed. See, e.g., Lambert 1993.) If we ignore the specific contents of sample records, then the distribution of this random variable depends on the population distribution only through its so-called *partition structure*, a term borrowed from the population genetics literature, where it has been widely studied. The partition structure is the distribution of the numbers of cells of each size.

For example, Chen and Keller-McNulty (1996) sample from an 87,959-element ''complete census from a single geographic region, taken during the 1980 decennial census'' (Zayatz 1991). In one cross-classification of this database according to five variables, there are 222 cells (Zayatz calls them ''equivalence classes'') of size one (the population uniques), 111 of size two, 73 of size three, etc. The largest cell-size is 3,649, and there is one such cell. In all, there are 1,024 non-empty cells. Chen and Keller-McNulty repeatedly sample from this cross-classified database, and compute, for each sample, the number of sample uniques and the number of these which are also population uniques. To perform this repeated sampling, they need only the partition structure, i.e., the vector

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_n) \tag{1}$$

where a_i is the number of cells of size i, $\sum a_i = K$, where K is the number of non-empty cells and $\sum i a_i = N$, where N is the population size. So, in the above example, $a_1 = 222$, $a_{3649} = 1$, $\sum a_i = 1{,}024$ and $\sum i a_i = 87{,}959$.

The above observation should not, however, raise false hopes. The mathematical problem of deriving, from a partition structure, the corresponding distribution of the proportion of sample uniques which are population uniques, is a formidable one, and attempting to solve it would almost surely be futile, because, undoubtedly, any expression which could be obtained would be quite intractable. (A simple, artificial – but highly illuminating – example of the problem is presented in Section 6.) A better approach is the Bayesian one, in which we treat the sample as a constant and the population as random – perhaps as itself a random sample from some *superpopulation*, which plays the role of a prior distribution. A favorite choice, in the literature on the uniques problem, has been the Poisson-Gamma model, by now largely discredited because of the poor performance of estimates derived from it; see, e.g., Skinner et al. (1994) or Keller and Bethlehem (1992). The model

I propose is a modification of the Poisson-Dirichlet model. (The name originates in Kingman (1975), though the model was studied earlier.) It comes from population genetics and species sampling, where it has been shown to play a central role in the study of partition structures. While researching this article, I was unaware that such ideas had ever been used in connection with the uniques problem for microdata; however, I subsequently learned of two other articles which use the Dirichlet-multinomial model to study other formulations of the uniques problem (see Omori 1998 and Takemura 1998).

To gain an understanding of genetic diversity – particularly in the so-called *neutral model* of evolution in the absence of selection (Kimura 1977) – population geneticists, and their allies in mathematics and statistics, have extensively studied the partition structure of the various *allelic states* of a gene, or of a number of genes. The partition structure of alleles is no different from that of cross-classified microdata files. Hence much of these researchers' work is immediately applicable to our problem. This includes the celebrated Ewens's sampling formula, originally presented in Ewens (1972), and subsequently shown to arise in numerous contexts. Also included is the Poisson-Dirichlet model – from which the Ewens's sampling formula arises naturally – and a two-parameter generalization described in Pitman 1996. In applying this work, it is best to begin with a simple model.

2. A Simplified Model

All of our models will be urn models. In this context, it is traditional, customary and convenient to speak, not of alleles or of cells arising from cross-classification of a microdata base, but of colored balls.

The setting is this: we have a population urn consisting of various numbers of various-colored balls. In the Chen and Keller-McNulty example cited above, there are 87,959 balls of 1,024 distinct colors, of which 222 colors are each represented by a single ball, while one color appears on 3,649 balls. In general, the composition of the population urn is unknown, but has a prior distribution which will be expressed by saying that the population is itself a random sample from some superpopulation of colored balls. We also have a sample urn, whose contents are, indeed, a random sample from the population (and, hence, a random subsample from the superpopulation). The way we make inferences about the population from the sample is to imagine sampling the remainder of the population, one by one, and invoking our prior distribution.

For example, suppose our prior distribution is completely *neutral*, i.e., is such that, throughout the process of sampling the remainder of the population, the distribution of the color of the next ball to be sampled is always simply the so-called *size-biased* distribution:

> The probability that the next ball to be sampled will be, say, Red, equals the proportion of Red balls presently in the sample.

Notice that if we do what is commonly called *Polya sampling*, namely draw a ball from the urn and replace it, together with another ball of the same color, the distribution of the new ball's color is exactly this size-biased distribution. Following custom, we call our sample urn, together with the size-biased sampling, a *Polya urn*.

Let the sample and population sizes be n and N, respectively. The uniques problem is then this: given a Polya urn whose initial state is our sample urn, and given $N - n$ draws,

what proportion of the colors which are unique in the initial urn state will remain unique at the final state? Conditioned on the initial state (so the denominator in the desired proportion is a constant), the numerator is the sum of the indicator functions of events of the form ''no new balls of this color are added.'' Each such event has probability $(n-1)/(N-1)$. So the mean of the proportion is just $[u(n-1)/(N-1)]/u$, where u is the number of uniques in the sample. The u's cancel, leaving, essentially, just the ratio of the sample size to the population size.

[The variance is also easy to compute – as a sum of variances of indicators plus a sum of covariances – and the result is also quite simple. It is approximately $(n/N)\{1-(n/N)(1-n/N)(u/n)\}$ which has negligible dependence on u when n/N is small.]

Is there such a prior distribution? From, e.g., Blackwell and MacQueen 1973 or Hoppe 1986 and Hoppe 1987, it would need to be such that the posterior distribution, given the sample, is the Dirichlet, $D(\alpha_1,\ldots,\alpha_k)$ distribution, where k is the number of colors (cells) in the sample, and the α_i's are the numbers of balls of each color in the sample. Such a prior can *almost* be achieved by using $D(\epsilon,\ldots,\epsilon)$ for very small ϵ. The superpopulation is then an infinite population in which the proportions of each color are random variables whose joint distribution is this Dirichlet prior.

So the model exists. But it is far from adequate, for at least two reasons:

- Empirically, we know that the proportion of sample uniques which are population uniques is often substantially larger than the ratio of sample size to population size.
- Logically, the model is too simple-minded because it makes no allowance for the introduction of new colors. It assumes that the sample already contains all the colors present in the entire population; hence, in particular, all population uniques are sample uniques.

3. The Poisson-Dirichlet Model

Here is a way of making the urn model more realistic: start with θ black balls. Each time a ball is drawn out, if it is not black do as before: replace it and add another of the same color. If it is black, replace it and add a ball of a new color, not yet in the urn. This obviously overcomes the second weakness of the first model, in that new colors will be introduced. And the expected proportion of sample uniques which are population uniques is now increased (by similar calculations) to $(n+\theta-1)/(N+\theta-1)$, which is at least a step in the right direction.

The parameter, θ, will, of course, need to be specified.

This model has been much studied in population genetics. In particular, Hoppe (1984) showed that the distribution of the partition structure, say Π_n, of the n non-black balls, after n draws, is given by the Ewens's (1972) sampling formula:

$$P(\Pi_n = \boldsymbol{a}) = \frac{n!}{[\theta]^n}\prod_{i=1}^{n}\frac{\theta^{a_i}}{i^{a_i}a_i!} \tag{2}$$

where $[\theta]^n = \theta(\theta+1)\cdots(\theta+n-1)$ and $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ is as in (1).

Furthermore, if, starting with just the θ black balls, we sample *infinitely many times*, and look at the *proportions* rather than the *numbers* of each color in the urn, and, at each stage, arrange the proportions in decreasing order, then there is a limiting distribution, called the

Poisson-Dirichlet distribution with parameter θ. It works this way. An infinite superpopulation is constructed, color by color. The proportion of the population with the first color has a beta distribution, with parameters 1 and θ. Then, the proportion of the remainder of the population with the second color also has a beta $(1, \theta)$ distribution, independent of the first. And so on: at each stage, the proportion of the remainder of the population with the next color is another I.I.D. beta $(1, \theta)$ random variable. It has been shown that if we now look backward at the sequence of urn drawings which created the superpopulation, then, conditioned on the values of all these infinitely many proportions, the colors of those successive balls which were added to the urn are I.I.D. from the superpopulation. See e.g., Hoppe 1987 or Kingman 1980. This is a very important result, the net effect of which is that a Poisson-Dirichlet prior distribution/superpopulation implies the validity of this Polya urn model.

What should θ be in this model? Clearly θ is related to the number of colors (classes) in the population. If the population size is N, then the expected number of distinct colors in the population is

$$1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+N-1} \approx \theta \ln\left(1 + \frac{N}{\theta}\right) \tag{3}$$

Hence, prior knowledge of the number of cells could be used to put a prior distribution on θ. Alternatively, one may estimate θ from the sample alone. From the Ewens's sampling formula (2), the maximum likelihood estimate of θ is easily seen to be the solution of

$$k = 1 + \frac{\theta}{\theta+1} + \frac{\theta}{\theta+2} + \cdots + \frac{\theta}{\theta+n-1} \approx \theta \ln\left(1 + \frac{n}{\theta}\right) \tag{4}$$

where k and n are, respectively, the number of distinct colors in the sample and the sample size.

My experience trying this model with the Chen and Keller-McNulty data has persuaded me that – at least with a constant θ – it is not sufficiently flexible. The problem is that the MLE of θ is too small, especially for relatively large sampling fractions. Rather than try mixtures (i.e., put a prior distribution on θ), I have opted for another parameter, because I believe this approach is easier to understand and work with. The two-parameter model described in Pitman 1996 is tempting because it includes a generalization of the Ewens formula, which allows the user to compute maximum likelihood estimates. But a closer examination of this model shows that it, in effect, imposes an even smaller estimate of θ. This is confirmed by the maximum likelihood estimates for my datasets. The model I use in the next section provides a much better fit to the data, but it sacrifices mathematical elegance (specifically, it lacks *exchangeability* which makes it much more intractable, mathematically).

4. An *ad hoc* Two-Parameter Model

In addition to the θ black balls, we start with M colored balls. These may be of various colors (let us call them primary colors) and are intended to skew the partition distribution by giving these colors a head start, thereby insuring that the population will have some very large cells. At each stage, if one of these M balls is drawn, it will, of course, be replaced, along with another of the same color. However these M balls – like the θ black

balls – do not count as part of the population or the sample. Since we are concentrating on sample and population uniques, and we presume that each of the colors among these M balls will be far from unique, it is irrelevant just how the M are partitioned among colors.

Now the expected proportion, p, of sample uniques which are population uniques gets bumped up still further, to

$$E(p) = (n + M + \theta - 1)/(N + M + \theta - 1) \tag{5}$$

and the expected number of new colors in the population becomes

$$\frac{\theta}{\theta + M} + \frac{\theta}{\theta + M + 1} + \cdots + \frac{\theta}{\theta + M + N - 1} \approx \theta \ln\left(1 + \frac{N}{M + \theta}\right) \tag{6}$$

A task for future research is to derive generalized Ewens and Poisson-Dirichlet distributions for this model.

What are reasonable guesses for M and θ? The quantity

$$L = \frac{M}{M + \theta} \tag{7}$$

is the expected proportion of primary-colored balls in both sample and population. So it should be roughly the proportion of records which are in ''large'' cells. But how large is large? To begin with, we should ask ourselves how well (5) can possibly work as an estimator of p if we are granted knowledge of how many population uniques are in the given samples and can choose $M + \theta$ opportunistically. If it fails to perform well even in this situation, there is no hope for it in the more realistic setting where we have only sample information (plus possible prior information for a Bayesian analysis). For the Chen and Keller-McNulty cross-classified data, cited in Section 1 of this article, I have opportunistically chosen L to be about .75. Substituting (7) into the right side of (6), and setting it equal to K, the number of cells in the population, gives

$$\theta \ln\left(1 + \frac{N(1 - L)}{\theta}\right) = K \tag{8}$$

Substituting, $L = .75$, rounding off N and K to be, respectively, $N = 88,000$ and $K = 1,000$, and solving for θ, we get, roughly, $\theta = 220$, and hence

$$M + \theta = \theta/(1 - L) = 880 \tag{9}$$

For each of the sample sizes, n, equal to 1%, 5%, 10%, and 50% of N, ten random samples have been generated, and the numbers of sample and population uniques have been computed. This allows us to compare the actual proportions of sample uniques which are population uniques with the predicted values from (5). Here are the results:

Table 1. Results for example 1

n/N	Predicted	Observed
.01	.020	.020
.05	.059	.074
.10	.109	.117
.50	.505	.514

(The observed values are the ratios of the total numbers of population uniques to sample uniques in all ten samples.) These very good results are not too sensitive to the choice of L. (Virtually the same results are obtained for $L = .70$ – in which case θ is about 210 and $M + \theta = 700$ – and for $L = .775$ – in which case θ is about 225 and $M + \theta = 1,000$. Indeed, results do not vary much over the whole range: $M + \theta = 900 \pm 300$.)

A slightly less rosy picture emerges from a second cross-classified dataset, from the same 89,000 record population in Chen and Keller-McNulty 1996. This one uses seven variables and has about 6,500 non-empty cells, of which about 3,100 are uniques. The largest cell-size is 2,378. I have again opportunistically chosen L to be .75. Using $K = 6{,}500$ in (8), gives θ of about 3,200 and, from (9), $M + \theta = 12{,}800$. Again ten random samples have been generated at each of the same sample sizes as before, and we combine sample and population uniques from all ten samples to get our observed values, and use (5) to get our predicted values. Here are the results:

Table 2. Results for example 2

n/N	Predicted	Observed
.01	.135	.100
.05	.171	.188
.10	.214	.273
.50	.563	.675

5. Estimation Using Sample Partition Structure

The good news in the previous section inspires us to go on to the more realistic case where we have only sample information. Suppose we have just one sample from a microdata population. Here is an *ad hoc* five-step procedure:

Step 1. Compute $\hat{\theta}_1$, the MLE from Model 2 (4); i.e., solve for θ in

$$\theta \ln\left(1 + \frac{n}{\theta}\right) = k \tag{10}$$

Step 2. Compute the cell-size threshold value, $e_n = 1 + n/\hat{\theta}_1$
Step 3. Let $1 - \ell$ be the proportion of sample records in cells of size $\leq e_n$
Step 4. Recompute $\hat{\theta}$ as the solution to

$$\theta \ln\left(1 + \frac{n(1-\ell)}{\theta}\right) = k \tag{11}$$

Step 5. Using $\hat{\theta}$ and $1 - \ell$ in (9) and (5), estimate the proportion of sample uniques which are also population uniques by

$$\hat{p} = \frac{n + \dfrac{\hat{\theta}}{(1-\ell)}}{N + \dfrac{\hat{\theta}}{(1-\ell)}} \tag{12}$$

Here are some comments on the above procedure:

1. The idea behind e_n, in Step 2, is this. In n draws from the Model 2 urn, the first color entered has, of course, the largest expected size of any color. That expected value is far too intractable to derive analytically, but it can be fairly well approximated by $1 + n/\theta$. We arbitrarily declare that any cell size larger than this approximate mean is ''Large.'' (Abstractly it would seem as if our threshold is on the small side, but when applied to the datasets, it tends to be if anything too large.)
2. In order to implement Step 3, I needed to generate my own samples because the data provided by Chen and Keller-McNulty does not include the sample partition structure need for Step 3. (Also, Laura Zayatz provided me with four additional population partition structures.) I used S-Plus software to generate and analyze numerous samples.
3. If we had a generalized Ewens sampling formula for Model 3, then we could replace Steps 1 to 4 by the maximum likelihood estimate. This is another subject for future research. A better approach would be to modify Model 3 in a way which retains its ease of use while increasing its mathematical tractability.
4. Notice that the number of sample uniques is not explicitly used at all in this procedure. (It enters implicitly when $e_n < 2$, in which case $1 - \ell$ is the proportion of sample uniques in the sample.) This is puzzling because obviously the number of population uniques in the sample is not independent of the total number of sample uniques. And yet, in the samples I have generated, the lack of dependence is quite striking (see Table 3). Undoubtedly there is a good theorem lurking here, just waiting for the right formulation.
5. If we have several samples, all of the same sample size, from a microdata population, we can combine the output to get a single $\hat{\theta}/(1 - \ell)$ in a natural way. If we have samples at various sample sizes (e.g., for $n/N = .01$, .05, .10, and .50), then how do we deal with the various $\hat{\theta}/(1 - \ell)$'s? One approach is simply to use the one from the largest sample size (which, in practice, seems also to be the largest value) because it is based on the most data. Of course, from (12), the larger $\hat{\theta}/(1 - \ell)$ is, the larger is $\hat{p}$.

Table 3. Summary of output for example 1

$n = 880$			$n = 4{,}400$			$n = 8{,}800$			$n = 44{,}000$		
up	*us*	*k*	*up*	*us*	*k*	*up*	*us*	*k*	*up*	*us*	*k*
5	104	205	11	157	413	20	175	541	105	212	866
0	94	197	12	171	418	25	198	556	118	213	881
5	110	206	11	169	426	19	182	537	115	212	878
2	99	201	8	161	413	15	173	518	124	231	884
1	88	196	9	166	422	15	179	540	115	234	878
1	111	205	10	150	405	26	193	544	106	221	870
3	112	214	8	149	401	19	175	529	100	214	865
3	110	205	10	171	424	22	193	551	108	215	863
4	107	208	8	158	418	21	190	537	113	207	867
1	102	200	11	159	411	28	200	559	106	208	872
25	1,037	2,037	98	1,611	4,151	210	1,858	5,412	1,110	2,167	8,724

Table 4. Five-step procedure: Results for example 1

n	$\hat{\theta}_1$	e_n	$(1-\ell)$	$\hat{\theta}/(1-\ell)$	Predicted $\hat{p}$	Observed $\bar{p}$
880	85	11	.49	300	.013	.024
4,400	110	41	.42	400	.054	.061
8,800	125	70	.37	500	.105	.113
44,000	155	285	.36	550	.553	.512

I have applied the five-step procedure to samples I generated from Chen and Keller-McNulty's Example 1 microdata population. As they did, I generated ten samples at each of the sample sizes n, equal to 1%, 5%, 10%, and 50% of N. Table 3 is a summary of the output (n is the sample size, *up*, *us* and k are, respectively, the numbers of population uniques, sample uniques and cells in the sample):

For Step 1, I used the average of the ten k's for each sample size. For Step 3, I used only the first of the ten samples. Table 4 shows what I obtained. The last column gives $\bar{p}$, the overall proportion of sample uniques which are population uniques in the ten samples being analyzed.

If I now follow my own advice about using the last value of $\hat{\theta}/(1-\ell)$, namely 550, at all four sample sizes, then the first three predicted values increase to .016, .056, and .106, respectively. (Recall that I used 880 when I could act opportunistically.)

The case $n = 880$ (i.e., the 1% sample) deserves special mention because the relative underestimate of $\bar{p}$ by $\hat{p}$ is substantial. With 220 population uniques among 88,000 records, the number of population uniques in a sample is nearly Poisson with parameter $\lambda = 880(220/88,000) = 2.2$. So the mean and standard deviation are 2.2 and $\sqrt{2.2} = 1.5$, respectively. Since there are roughly 100 sample uniques in each sample, the ratio, *up*/*us*, varies quite a bit; e.g., $(2.2 - 1.5)/100 = .007$, while $(2.2 + 1.5)/100 = .037$. So, obviously, a practical method needs much more than just the point estimates of our *ad hoc* procedure.

I have also analyzed the other population partition structures with which I was provided. They all have substantially higher proportions of population uniques, which makes them less interesting because, even with 1% sampling, *up*/*us*, as well as our estimate of it, is so large that disclosure risk is unacceptably high. For example, for a population with 50,000 records, of which 8,160 are uniques, ten samples of size $n = 500$ looked as follows:

Table 5. Summary of output for another example

up	89	88	81	67	86	84	71	81	86	66	799
us	427	431	429	425	429	426	423	425	428	419	4,262
k	457	459	459	459	461	455	455	461	460	455	4,581

(same notation as in Table 3)

Again using the average of the k's (which are remarkably stable), we get $\hat{\theta}_1 = 2,700$, $e_n < 2$. If we accordingly call all but the sample uniques ''large'', we have, in the first sample, $1 - \ell = .854$. But solving for $\hat{\theta}/(1-\ell)$ is equivalent to solving

$$x \ln\left(1 + \frac{n}{x}\right) = \frac{k}{1-\ell}$$

But with $k = 457$ and $1 - \ell = .854$, there is no solution because the ratio on the right side is larger than 500, while the left side is an increasing function of x with limit $n = 500$ as $x \uparrow \infty$. So our *ad hoc* procedure forces us to say $\hat{\theta}/(1 - \ell) = \infty$, which is equivalent to saying $\hat{p} = 1$. The result is that we are acting very conservatively, since the true value (overall for the ten samples) is $799/4,262 = .19$.

6. A Cautionary Note

If the preceding result seems unsatisfactory, it would be well to consider the challenges presented by the data set. The sample contains 500 records in 457 cells, of which only 30 have more than one record. Indeed, there is one cell of size 5, three of size 4, four of size 3 and 22 of size 2. With such partition structure, I claim that no estimate, based only on the data, can have much predictive value (in which case one needs to be very conservative). Here is an example to illustrate the point.

Consider two populations, both consisting of N records. Population 1 has all uniques, while Population 2 has all cells of size 2, hence no uniques. Now take random samples of size n from each population. The Population 1 sample will, of course, consist entirely of sample uniques. In Population 2, each cell has probability nearly $(n/N)^2$ of having both of its records in the sample. There are $N/2$ cells in the population, so the expected number of sample cells with two records is $N(n/N)^2/2$, which is $(n/2)(n/N)$, and the expected number of sample records which are not sample uniques is twice that, or $(n/N)n$. Thus, if $n/N = .01$, we expect 99% of all sample records from Population 2 to be sample uniques, vs l00% from Population 1. With samples from the two populations looking so similar, how can we possibly hope to have an estimate, $\hat{p}$, close to one for the first sample, yet close to zero, for the second one? To be safe, we must opt for the former at the expense of the latter.

7. Future Work

The work described in this article is obviously just a beginning. It explores the possibilities of adapting ideas and models borrowed from the mathematical theory of population genetics to provide disclosure risk assessment for microdata. I think it clearly makes a good case for further exploration. In its present state, the methodology described here does not take into account the fact that partition structure results from cross-classification. Nor does it – as has been pointed out – yet address itself adequately to the variability inherent in the sampling process. As we learn more about the processes that produce partition structures, we should be able to exploit the mathematical ideas of majorization and convexity to partially order these structures in terms of inherent disclosure risk.

8. References

Blackwell, D. and MacQueen, J.B. (1973). Ferguson Distributions via Polya Urn Schemes. Annals of Statistics, 1, 353–355.

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

de Waal, A.G. and Willenborg, L.C.R.J. (1996). A View On Statistical Disclosure for Microdata. Survey Methodology, 22, 95–103.

Duncan, G.T., Jabine, T.B., and de Wolf, V.A., (eds.) (1993). Private Lives and Public Policies. Confidentiality and Accessibility of Government Statistics. Washington: National Academy Press.

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

Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data. Journal of Official Statistics, 14, 385–397.

Hoppe, F.M. (1984). Polya-like Urns and the Ewens' Sampling Formula. Journal of Mathematical Biology, 20, 91–94.

Hoppe, F.M. (1986). Size-biased Filtering of Poisson-Dirichlet Samples with an Application to Partition Structures in Genetics. Journal of Applied Probability, 23, 1008–1012.

Hoppe, F.M. (1987). The Sampling Theory of Neutral Alleles and an Urn Model in Population Genetics. Journal of Mathematical Biology, 25, 123–159.

Keller, W.J. and Bethlehem, J.G. (1992). Disclosure Protection of Microdata: Problems and Solutions. Statistica Neerlandica, 46, 5–19.

Kimura, M. (1977). The Neutral Theory of Molecular Evolution and Polymorphism. Scientia, 112, 687–707.

Kingman, J.F.C. (1975). Random Discrete Distributions. Journal of The Royal Statistical Society, Series B, 37, 1–22.

Kingman, J.F.C. (1980). Mathematics of Genetic Diversity. CBMS–NSF Regional Conference Series. In Applied Mathematics, 34. S.I.A.M. Philadelphia.

Kirkendall, N.J., Arends, W.L., Cox, L.H., de Wolf, V., Gilbert, A., Jabine, T.B., Kollander, M., Marks, D.G., Nussbaum, B., and Zayatz, L.V. (1994). Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22. Federal Committee on Statistical Methodology. U.S. Office of Management and Budget.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–331.

Omori, Y. (1998). Measuring Identification Disclosure Risk for Categorical Microdata by Posterior Population Uniqueness. (Preprint presented at SDP'98, Lisbon, Portugal, 25 March.)

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. Statistics, Probability and Game Theory. IMS Lecture Notes-Monograph Series, 30, 245–267.

Skinner, C.J., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure Control for Census Microdata. Journal of Official Statistics, 10, 31–51.

Takemura, A. (1998). Some Superpopulation Models for Estimating the Number of Population Uniques. (Preprint presented at SDP'98, Lisbon, Portugal, 25 March.)

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

Received September 1997
Revised May 1998