WP.27
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

THE APPLICATION OF THE CONCEPT OF UNIQUENESS FOR CREATING PUBLIC USE MICRODATA FILES

Invited Paper

Prepared by Jay J. Kim (National Center for Health Statistics, United States of America)[1] and Dong M. Jeong (Korea National Statistical Office, Republic of Korea)

[1] The findings and conclusions in this paper are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

The Application of the Concept of Uniqueness for Creating Public Use Microdata Files

Jay J. Kim [1,3] and Dong M. Jeong[2]

[1] National Center for Health Statistics, Hyattsville, MD 20872, USA
[2] Korea National Statistical Office, Daejeon, Republic of Korea

Abstract. In general, agencies use non-systematic ad hoc approaches for protecting against disclosure in microdata. This paper develops probability models quantifying disclosure risk for a microdata file. This is a modification of the Marsh, et al (1991) procedure. The model can use population and sample uniques only, or it can also include population and sample twins or triplets. For identifying population uniques, twins or triplets, we need to determine what type of information intruders have which is also available on the microdata file. This common information is called "key variables." Using the models, disclosure risk can be computed for the original microdata, and we can determine whether the risk is too high or not. If the risk is too high, grouping categories, post randomization methods or randomized response techniques, etc., can be used for masking the categorical variables. If the variable is continuous, grouping or noise inoculation, etc., can be used for masking the variable. The probability model using population and sample uniques only was applied for creating disclosure-limited microdata files using the 2005 Korean demographic census data (8). In this paper, an attempt is made to develop a theoretically defensible and systematic approach for protecting against disclosure in microdata.

Keywords. disclosure risk, population and sample uniques, key variables, grouping

1 Introduction

Government agencies release microdata files from their survey data or administrative records data. Large amounts of information on individuals is available to many organizations and data users. If a public use microdata file (PUMF) is released, intruders could try to match their records with the ones from the PUMF and gain access to new information. Note that in linking the records on two files, common information is used, which is called "key variables." Data disseminating agencies must protect the confidentiality of individuals on their files. In the U.S., laws such as Title 13 stipulate protection of the confidentiality of many types of data such as survey data and income tax return data. On the other hand, agencies should not ignore the data users' need, i.e., the utility of the data files. Therefore, in creating a

[3] The findings and conclusions in this paper are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

PUMF, agencies attempt to eliminate disclosure risk of the file while maintaining maximum utility of the data.

PUMF does not carry names and addresses of individuals on the file. Thus intruders have to rely on other variables. However, even if records on two files match exactly, there is no guarantee that they represent the same individual(s). If they are population and sample uniques, no errors are in the data (key variables), and the data are collected almost at the same time (the reference periods are similar), the match would be correct, but if an individual is not a population unique, there is still a chance of identification, but the probability would be lower.

Marsh, et al (1991) developed a probability model for measuring disclosure risk for creating a PUMF from the United Kingdom's census data. Their model depends on population uniques. In this paper we will show modified versions of the model and apply them for creating a PUMF using the 2005 Korean census data.

2 Intruders and Disclosure

Potential intruders may attempt to match the records on the PUMF to their databases which contain identifiers and glean new information from the PUMF. Examples of potential intruders are credit card companies, mortgage departments of banks, insurance companies, credit bureaus, trade associations, central government agencies such as Internal Revenue Service and local government agencies which have access to organizational data base. However, intruders do not need to be organizations holding data sets. There is a lot of information in the public domain and because of readily available high computing power, individual intruders can assemble data files themselves.

There are two types of disclosure. The first is an identity disclosure. Lambert (1993) calls this identification. If the intruder is a journalist and tries to embarrass the data disseminating agencies, his claim that he has been successful in identifying someone on their PUMF would be sufficient. If the intruder publicizes the findings in the news media, it could have a devastating effect on the agencies' data collection efforts. The other is gaining new information after identification (2, 9). Lambert calls this an attribute disclosure. Identity disclosure is a prerequisite for disclosure of additional information. Variables other than key variables are available in the PUMF, and hence once the identity is disclosed, there is no doubt that new information can be disclosed. Thus for defining a measure of disclosure risk, identity disclosure and attribute disclosure will be considered equivalent in this paper.

3 Measures of Disclosure Risk

Let

P(a) = the probability of key variables being recorded identically in both PUMF and intruder's file;

P(b|a) = the probability that an individual appears in a PUMF is the same as the sampling fraction for that individual in the PUMF;

P(c|a,b) = the probability of population unique;

and

P(d|a,b,c) = the probability of verifying population uniqueness.

Marsh, et al (1991) defined the probability of correct identification of an individual as

$$P(\text{correct identification of an individual}) = P(a)P(b \mid a)P(c \mid a,b)P(d \mid a,b,c) \quad (1)$$

We modify the above model in the following manner.

Disclosure can occur whether a person in a population is unique or not based on the values of key variables. That is, even if there are two or three persons in a population who are the same on key variables, one of their identities could be disclosed, if additional information is accessed, because more information than that on key variables is available on the PUMF and intruder's file. However, most researchers have been paying attention to the population unique cases only in defining the disclosure risk. Thus, the disclosure risk can be defined narrowly or broadly depending on whether or not it is restricted to the unique cases in the population. Here, we will develop formulas for both narrow and broad definitions of the disclosure risk. In deriving the formulas, we assume that there are no recording or classification errors for the values of the key variables. In other words, *P(a)* = 1 in Marsh, et al's formula. It is also assumed that *P(d|a,b,c)* = 1. Our model will be restricted to the sample unique.

Disclosure can occur when the following conditions are met:

1. An individual is unique in a population based on key variables.
   If the intruder's file is a 100 percent population file, he can establish uniqueness of a certain individual by using his file.
2. The individual is on a PUMF from a survey or an administrative records file.
3. The individual is on another file which an intruder is working with.
   An intruder can have information on key variables for a specific person and try to examine whether that person appears in the above microdata. In this case, the second file has a single record.
4. The individual on the files in conditions 2 and 3 above is unique.

In general, agencies use non-systematic ad hoc approaches for protecting against disclosure in microdata. It is well known that the disclosure risk is affected by the inclusion probability in the file(s) (3). In this paper we develop comprehensive probability models for the disclosure risk which incorporate the inclusion probability in the PUMF and intruder's file.

Let

A = an individual of interest;

F_1 = PUMF in condition 2 above

F_2 = a file held by an intruder (in condition 3 above);

P_1 = unique class in the population;

S_{1F_1} = unique class in F_1;

and

S_{1F_2} = unique class in F_2.

3.1 A Narrow Definition of Disclosure Risk

The narrow definition of disclosure risk is based on the population and sample uniques only.

3.1.1 Assuming an Intruder Does a Phishing (Fishing) Expedition

Here we assume that the intruder does not know anyone on PUMF, but tries to link the records on his file to those on the PUMF. When all four conditions mentioned above and the assumption of no measurement error are met, then the probability of correct identification is

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \cap \left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \cap \left(A \in P_1\right)\right] \tag{2}$$

If an individual is a population unique, it would also be a sample unique. In other words,

$$\begin{aligned} &P\left[\left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \cap \left(A \in P_1\right)\right] \\ &\quad = P\left[\left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \mid \left(A \in P_1\right)\right] P\left(A \in P_1\right) \\ &\quad = P\left(A \in P_1\right) \end{aligned}$$

Because of the above, the disclosure risk for individual A can be reduced to the following.

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \cap \left(A \in P_1\right)\right] \tag{3}$$

which can be further re-expressed as follows:

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \mid \left(A \in P_1\right)\right] P\left(A \in P_1\right) \tag{4}$$

Note that the event that A is unique in population is independent of whether A is selected in sample or not. Thus, equation (4) reduces to

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right)\right] P\left(A \in P_1\right) \tag{5}$$

The event that A is in the PUMF is usually independent of the event that A is in the intruder's file. For example, the event that an individual is selected in a labor survey is independent of the event that it is on the income tax return file. In this case, equation (5) can be simplified as

$$P\left(A \in F_1\right) P\left(A \in F_2\right) P\left(A \in P_1\right) \tag{6}$$

However, a survey can be a subset of another survey. Thus if F_2 is a larger survey and F_1 is a subset of F_2, then

$$P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right)\right] = P\left(A \in F_2\right) P\left[\left(A \in F_1\right) | \left(A \in F_2\right)\right]$$

Note that $P\left[\left(A \in F_1\right) | \left(A \in F_2\right)\right]$ is a subsampling rate of F_1 from F_2. Thus if there is dependence between the two files, equation (6) becomes

$$P\left(A \in F_2\right) P\left(A \in P_1\right) \text{Subsampling Rate} \tag{7}$$

If F_2 is a 100 percent file and $P\left(A \in P_1\right)$ is calculated using F_2, S_{1F_2} in equation (2) should be ignored. S_{1F_2} is needed in equation (2) when F_2 is not a 100 percent census file.

3.1.2 Assuming an Intruder Already Knows That A is in F_1

If an intruder already knows that A is on the PUMF, that is, if the intruder has response knowledge (1), then $P\left(A \in F_1\right) = 1$. Thus, from equation (6), the disclosure risk will be

$$P\left(A \in F_2\right) P\left(A \in P_1\right) \tag{8}$$

Note that if the intruder's file is a 100 percent population file, then $P\left(A \in F_2\right) = 1$. The disclosure risk is

$$P\left(A \in P_1\right) \tag{9}$$

3.2 Broader Definition of Disclosure Risk

Even if an individual is not unique in the population, he/she still can be identified with additional information included in the PUMF and intruder's file. Thus, the definition of disclosure risk can be broadened. Suppose C individuals in the population have the same values of the key variables and matching to any one of them is equally likely. Define

P_C = Equivalence class of size C in the population.

Then, the probability of correct identification is

$$\frac{1}{C} P\left[\left(A \in F_1\right) \cap \left(A \in F_2\right) \cap \left(A \in S_{1F_1}\right) \cap \left(A \in S_{1F_2}\right) \cap \left(A \in P_C\right)\right] \tag{10}$$

In the above, since a sample unique individual can be matched to either one of the C individuals in population, a multiplier of 1/C is needed.

4 Evaluation of Disclosure Risk

In Korea, there exists a full national population register and every member of the nation excluding foreigners is assigned a resident registration number similar to Social Security Numbers in the U.S. On the register, a person's name, gender, birth year, month and date, place of birth and the place of registration are available. The Korean government does not make this information available to the public. However, commercial enterprises make this information readily available on the web. Because of this situation, one of the missions of the Korean National Statistical Office (KNSO) is to make their PUMF disclosure protected.

For this study, we selected the 2005 census data from a Korean province, Choongchung (CC) Province. The population size, and the number of households and housing units are shown in Table 1. Note that a family or household can be disclosed first, then its members can be further identified. In this paper, however, we will direct our attention to the direct disclosure. The 2 percent microdata file is created by taking a 20 percent subsample of the 10 percent census sample. Thus, the 2 percent microdata corresponds to F_1 and the census sample to F_2. The probability of a population unique is calculated using the 100 percent census.

Table 1. Population Size, and Number of Households and Housing Units - CC Province

	Population	Households	Housing Units
Census	1,798,397	660,526	586,757
Census Sample (10%)	189,505	71,091	65,398
2% Microdata	38,027	14,218	13,038

The following are matching keys we used: gender (2); age (111); marital status (4); relationship to householder (14); household type (5); tenure (6); building type of residence (12); and type of housing and number of floors of the building (12). The number in the parentheses is the number of categories for each of the variable. All the key variables are discrete.

Using the census file (N = 1,798,397) with all eight variables, 9,664 persons were unique, which is 0.54 percent of total population. This is much higher than Bethlehem, et al's threshold of 0.1 percent. If we assume that the intruder has a 10 percent census sample file, the disclosure risk is $0.1 \times 0.2 \times 0.0054 \approx 0.00011$, according to equation (7). However, whole blocks are selected in the 10 percent census sample, thus residents in the sample blocks know that their neighbors are also in the sample, i.e., many sample persons have response knowledge. To those who have response knowledge, the disclosure risk is $0.2 \times 0.0055 \approx 0.0011$, according to equation (8). If we assume that the intruder has the 100 percent population file, the

risk is 0.54 percent. The risks are too high, thus we decided to coarse categories of selected variables. This approach is often used to reduce the risk (1, 5, 9).

Table 2 below shows how the grouping affects the probability of uniques by using up to 4 variables. In the table, x in the columns 2 – 5 indicates which variable(s) is (are) used for identifying uniques.

Table 2. Number of Unique Persons before Grouping Categories –Population Data

# of Vars	Gender	Age	Relationship	Marital Status	# of Uniques
1	x				0
1		x			2
1			x		0
1				x	0
2	x	x			5
2	x		x		0
2	x			x	0
2		x	x		65
2		x		x	11
2			x	x	0
3	x	x	x		167
3	x	x		x	30
3	x		x	x	2
3		x	x	x	349
4	x	x	x	x	713

Table 2 above shows that as we use more variables for identification, the number of uniques increases. Note that the number of uniques with more variables also includes the number of uniques with fewer variables. Note also that the variable age is in single years. When only one key variable is used, age alone provides 2 unique ones. When gender and age are used, the total number of uniques increases to 5. These 5 uniques includes the 2 uniques due to age alone. That is, the unique counts are cumulative. When age and relationship are used, the number of uniques becomes 65. When age, relationship and marital status are used, the total number of unique persons is 349. When marital status is included along with gender, age and relationship, the number of uniques increases to 713.

Table 3 below shows the number of uniques before and after age groupings. Before the grouping, there were 111 age categories, but after the grouping, the number of age categories was reduced to 20.

Table 3. Number of Uniques with 5 Year Intervals for Age – Population Data

# of Vars	Gender	Grouped Age	Relationship	Marital Status	# of Uniques

1		x			2 → 0
2	x	x			5 → 2
2		x	x		65 → 6
2		x		x	11 → 1
3	x	x	x		167 → 18
3	x	x		x	30 → 3
3		x	x	x	349 → 53
4	x	x	x	x	713 → 106

In Table 3 above, the last column shows how much the number of uniques in Table 2 gets reduced due to age groupings. The first number in the column is the number of uniques before grouping and the latter is the number after grouping. Note that by grouping age alone into 5 year intervals, the number of uniques was lowered, sometimes by the ratio of 11:1. If 10 years of age are grouped, with 4 variables, the number of uniques becomes 51, less than half of the 106 uniques when age was grouped into 5 year intervals.

Comparing Table 4 below with Table 3, in Table 4, the variable relationship is also grouped.

Table 4. Number of Uniques with Grouped Age and Relationship Categories – Population Data

# of Vars	Gender	Grouped Age	G. Relationship	Marital Status	# of Uniques
2	x	x			2 → 2
2		x	x		6 → 2
2		x		x	1 → 1
3	x	x	x		18 → 4
3	x	x		x	3 → 3
3		x	x	x	53 → 3
4	x	x	x	x	106 → 8

Table 4 above shows that by using grouped relationship, the number of uniques gets further reduced most of the time.

In comparison to Table 4 above, Table 5 below has grouped marital status.

Table 5. Number of Uniques with Grouped Age, Relationship and Marital Status Categories – Population Data

# of Vars	Gender	G. Age	G. Relationship	G. Marital Status	# of Uniques
2	x	x			2 → 2
2		x	x		2 → 2
3	x	x	x		4 → 4

3	x	x		x	3 → 1
3		x	x	x	3 → 1
4	x	x	x	x	8 → 4

Table 5 shows that grouping marital status sometimes reduces the number of uniques, but not as much as grouping by age or relationship.

We tried two different groupings in terms of the number of categories for: (i) relationship, (ii) building type, and (iii) type of housing and the number of floors of the building. The first grouping has 9, 6, and 6 categories in order (as previously mentioned) and the other grouping has 3, 4 and 4 categories. The first grouping provides 501 uniques and the second results in 495 uniques. The difference is minor. Both represent around 0.028 percent of the total population. This is much lower than Bethlehem, et al's threshold of 0.1 percent. If we assume the intruder has the 10 percent census sample file, the disclosure risk is 0.0000056. This translates into one person per 100,000 people. This is a very low risk. Some agencies provide geographical information if an area has at least 100,000 people. If we assume response knowledge, the disclosure risk goes up to 0.000028. However, if we assume that the intruder has the 100 percent population data, then the risk further goes up to 0.028 percent. This means 28 persons per 100,000 people. In addition to grouping categories of key variables, other variables such as occupation, lot size, and the size of living space need to be investigated.

5 Concluding Remarks

Comprehensive probability models quantifying disclosure risk have been developed for microdata files. The measure of disclosure risk can depend on the number of population uniques, population twins or triplets. We developed two probability models, one for the population unique case and the other for population twins or triplets, etc. The models assume that only unique persons in the sample files face disclosure risk.

We measured the probability of the population uniques using the original census data of KNSO which was 0.54 percent. Assuming that an intruder performs a fishing expedition using the sample data from the census, disclosure risk will be 0.011 percent. If the intruder has response knowledge, the disclosure risk goes up to 0.11 percent. If the intruder has the 100 percent census data, the risk goes up further to 0.54 percent. In this case, the threshold of Bethlehem, et al (1990) is exceeded.

To lower the probability of identifying uniques, we grouped categories by 5 year age intervals. Two different numbers of categories were used for three variables (relationship, building type, and type of housing and the number of floors of the

building), one grouping has fewer categories in all three than the other grouping. However, both provided similar probability of identifying uniques, 0.028 percent, which is much lower than Bethlehem, et al's threshold. In this case, if we assume the intruder has the 10 percent census sample file, the disclosure risk is 0.0000056. This means one person per 100,000 people. Even if we assume response knowledge, the disclosure risk is 0.000028. Thus, the above grouping seems to provide sufficient disclosure protection.

Using this approach, the second author created PUMFs for KNSO which were released to the public. It should be noted that, in addition to the above, the second author used the broadest occupation codes and masked some other variables such as the size of living space and lot size.

References

1. Bethlehem, J., Keller, W., and Pannekoer, J. (1990). *Disclosure Control for Microdata, Journal of the American Statistical Association*, 85, 38-45.
2. Cox, L.H. and Sande, G. (1979). *Techniques for Preserving Statistical Confidentiality, Bulletin of the International Statistical Institute*, 42.3, 409-512.
3. Elliot, M. (2001), *Disclosure Risk Assessment, in: Confidentiality Disclosure and Data Access, North-Holland* 75-90.
4. Feinberg, S. and Markov, U. (1998), *Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data, Journal of Official Statistics*, 14, 385-397.
5. Khare, M., Battaglia, M.P. and Hoaglin, D.C. (2003). *Procedures to Reduce the Risk of Respondent Disclosure in a Public-Use Data File: The National Immunization Survey, Federal Committee Statistical Methodology Research Conference*, 5-12.
6. Lambert, D. (1993), *Measures of Disclosure Risk and Harm, Journal of Official Statistics*, 9, 313-331.
7. Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991), *The Case for Samples of Anonymized Records from the 1991 Census, Journal of the Royal Statistical Society A, 154, Part 2*, 305-340.
8. Korean population census, www.nso.go.kr/eng2006/emain/index.html
9. Skinner, C., Marsh, C., Openshaw, S., and Wymer, C. (1994), *Disclosure Control for Census Microdata, Journal of Official Statistics, 10*, 31-51.