IAOS Satellite Meeting on Statistics
for the Information Society
August 30 and 31, 2001, Tokyo, Japan

On Recent Developments in Statistical Disclosure Control Techniques

TAKEMURA, Akimichi
Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo, JAPAN

Abstract

Disclosure control of microdata sets is an important practical topic for statistical agencies. It is also a very interesting problem for theoretical statisticians from the viewpoint of statistical inference. In recent years some significant theoretical developments have been achieved by a group of Japanese researchers including the author. Here we summarize some of our results.

1 Introduction

With the rapid development of personal computers and Internet, there is a growing demand for more microdata sets from the users of statistical data. With popular software for data analysis, the users themselves can easily handle microdata sets with their personal computers. In the U.S. and some countries in Europe, it is now common that the statistical offices publish microdata sets in addition to summary tables, if the disclosure risk of a microdata set is considered to be sufficiently low. Unfortunately in Japan virtually no microdata set is published by the government. This is partially due to their very strict interpretation of Japanese statistics law. Another factor is that disclosure control procedures are not yet sufficiently standardized.

Disclosure risk of a microdata set depends on many factors. It may be the case that the disclosure risk a microdata set is low, even if the (re)identification risk of individuals in the data set is relatively high, simply because the information contained in the microdata set is not "interesting" from the viewpoint of an attacker. Therefore formal mathematical evaluation of identification may not be sufficient for the overall evaluation of the disclosure risk. On the other hand the mathematical evaluation itself is important because of it gives an objective measure of the disclosure risk. This is similar to the fact that in sample surveys evaluation of sampling error is the basis for evaluation of their variability, although there are many other sources of variation and biases in actual sample surveys. With this fact in mind, we review some new results of the author and collaborators on the evaluation of identification risk and related SDC techniques. Comprehensive survey of SDC techniques is given in Willenborg and de Waal (2001).

2 Estimation of the number of population uniques in the sample

The most important measure of identification risk of a microdata set is the estimated number of population uniques among the sample uniques of the microdata set. From the viewpoint of statistical inference this problem is very difficult, because the number of small cells in the sample does not contain enough information on the number of small cells in the population under usual low sampling ratio (less than 1/1000). Therefore we need rather strong a priori assumption on the distribution of

number of small cells in the population.

One way of imposing reasonable assumption is the superpopulation model, where the actual population is regarded as generated from a hypothetical superpopulation. In the disclosure field, Bethlehem et al.(1990) proposed to use the Poisson-Gamma model. There have been many similar models employed in the field of statistical ecology and these models can be used in the disclosure field. In Takemura(1999) and Hoshino and Takemura(1998) we compared some well known models, including the multinomial-Dirichlet model, the Ewens sampling formula and the logarithmic series model of Fisher. We found that these models are closely related. Actually, application of these models to microdata sets often leads to similar estimates of the number of population uniques. Properties of the Poisson-Gamma model and the related models are now well understood.

A rather different estimate of the number of population uniques in the sample is obtained by the Pitman sampling formula (Hoshino(2000)). The Pitman sampling formula is a generalization of Ewens sampling formula and it seems to fit frequencies of the cell sizes of microdata sets much better than the Poisson-Gamma model. Compared to the Pitman sampling formula, the Poisson-Gamma model and closely related models seem to underestimate the number of population uniques in the sample. On the other hand the Pitman sampling formula seems to overestimate the number of population uniques, although at this point it is difficult to evaluate the performance of the Pitman sampling formula, because we (the author and collaborators) do not yet fully understand the nature of the Pitman sampling formula.

Omori(1999) employs a full Baysian approach to the problem and evaluate the number of population uniques in terms of the posterior probability of population uniqueness. Sai and Takemura (2000) used the Poisson-Gamma model to describe a simple procedure of evaluating the decrease in identification risk by further global recoding of categories in a key variable.

3 New methodology for local recoding and record swapping

Obviously the global recoding is the most important SDC measure. It is simple and its interpretation is clear both from the viewpoint of the user and of the provider of microdata sets. However some more fine-tuned SDC measures are often desirable, because microdata sets might lose too much information when the global recoding is too freely used. Most fine-tuned measures are perturbative ones including adding noise to quantitative variables and randomly changing categories for qualitative variables. PRAM (post randomization method) is currently actively studied for the random alteration of categories (see Chapter 5 of Willenborg and de Waal(2001)).

In Takemura (2001) the author proposed a technique for local recoding and data swapping by matching close records in a microdata sets into disjoint pairs. For optimal matching a well known algorithm of maximum weight matching ("Edmonds' algorithm") can be employed.

When the records are matched in pairs, the actual observations can be displayed as an interval or union of categories covering just these two records. This is a form of "local recoding", where recoding is done only for specific records and variables. The pairing can also be used for the purpose of record swapping, where observations of the two records are swapped if necessary.

We can interpret the local recoding as shifting the record swapping from the producer of the microdata sets to the users. Users themselves can randomly choose possible values from the locally recoded intervals or union of categories. The resulting data set is a particular data set obtained by the application of record swapping to the original microdata set. An added advantage in this shifting of the swapping from the producer to the users is that the users can themselves evaluate the added variability due to the local recoding. Despite this advantage of local recoding, the statistical agencies may still prefer record swapping because of its simplicity.

4 Evaluation of per-record identification risk

Overall evaluation of the identification risk can be achieved by the superpopulation model approach discussed in Section 2. However these models only give an overall estimate of the number of the population uniques in the sample. These models do not suggest which records have higher identification risk than others, because these models treat the record "exchangeably", i.e., in an interchangeable manner. In order to apply perturbative disclosure control measures of the last section, we have to evaluate the per-record identification risk of the records in a microdata set.

One descriptive way of evaluating which sample unique records are at particularly high identification risk is to count the smallest number of key variables, for which the record is already a sample unique. Sample unique records which are unique with only a few key variables are obviously at higher identification risk than sample unique records which are unique with respect to combinations of many key variables. This is the idea of the minimum unsafe combination of variables. In Takemura(1999) the author have discussed in detail the notion of the minimum unsafe combination of variables and its associated notion of maximum safe combination of variables.

Another general approach to the per-record identification risk is to categorize all the key variables, treat the data set as a multiway contingency table and fit a model to the contingency table. Then the sample unique cells with low estimated cell probability can be considered unsafe.

Fienberg and Makov(1998) and Skinner and Holmes(1998) fitted the standard log-linear model for this purpose. Recently, the author (Takemura(2001)) proposed to use additive cell probability model for contingency tables for its simplicity. Fitting of log-linear model to contingency tables with many (say 10) variables is computationally expensive. The additive probability model can be fitted easily to large contingency tables with many variables. In actual evaluation of identification risk, it is no uncommon to consider more than 10 key variables. Therefore additive probability model is advantageous for its computational simplicity. The disadvantage of the additive probability model is that it assumes no "structural zeros" (cells where there is no observation by definition) in the contingency tables. In actual microdata sets, there are many structural restrictions among the key variables and usually there are many structural zeros. The fit of the additive probability model is not necessarily good because of existence of many structural zeros.

REFERENCES

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the Americal Statistical Association*, **85**, 38-45.

Fienberg, S.E. and Makov, U.E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data. *Journal of Official Statistics*, **14**, 385-398.

Hoshino, N. (2000). Applying Pitman's sampling formula to microdata disclosure risk assessment. Submitted for publication.

Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation model useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, **28**, 125-134.

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.* Office for Official Publications of the European Communities, Luxembourg, 59-76.

Sai, S. and Takemura, A. (2000). Some models for merging groups in microdata. *Japanese Journal of Applied Statistics*. **29**, 63-82. (in Japanese).

Skinner, C.J. and Holmes, D.J. (1998). Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, **14**, 361-372.

Takemura, A. (1999). Calculating minimum k-unsafe and maximum k-safe sets of variables for disclosure risk assessment of individual records in a microdata set. ITME discussion paper No.6, Faculty of Economics, University of Tokyo.

Takemura, A. (2001). Evaluation of per-record identification risk by additive modeling of interaction for contingency table cell probabilities. To be presented at the invited paper session (IPM18) on Disclosure Control at the 53rd Session of ISI.

Takemua, A. (1999). Some superpopulation models for estimating the number of population uniques. *Statistical data protection - Proceedings of the conference, Lisbon, 25 to 27 March 1998 - 1999 edition.* Office for Official Publications of the European Communities, Luxembourg, 45-58.

Takemura, A. (2001). Local recoding and record and record swapping for by maximum weight matching of for disclosure control of microdata sets. *Journal of Official Statistics.* Conditionally accepted.

Willenborg L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer, New York.