Measures of Data Utility for Complex Survey Data

Kyu-Seong Kim *

Abstract

When statistical data are released to the public statistical disclosure control techniques may reduce the risk of disclosure of confidential information on respondents. On the other hand information loss of some interested variables is inevitable due to such techniques. The existing data utility measures in order to evaluate the information loss have been suggested mainly based on IID data, which means that each record is dealt with evenly. But in case of complex survey data since the importance of each record is different, each record should be dealt with differently even in data utility measure.

This paper focuses on data utility measure for complex survey data. Two data utility measures based on the weighted empirical distribution functions are proposed, where weighted empirical CDF is estimated from not only original and masking data but also survey weights attached to records. A simulation study conducted on 2006 Korea welfare panel data shows that the existing measures based on IID data are much lower than the proposed measures, which means that the existing measures may report data utility to be much more useful than actual usefulness in case of complex survey data.

Key Words: Original and protected data, survey weighted, weighted empirical CDF.

1. Introduction

The need for microdata by researchers has been increasing rapidly, because microdata, in reality, is a basic source of much information many researchers want to obtain. A lot of statistical disclosure control techniques have been developed in order to protect respondent's identity when microdata is released. Respondent's confidentiality can be partly protected by using such techniques, whereas we do look at the loss of information as well. In other words, the utility of microdata altered by some disclosure control techniques will be decreasing with greater use of such techniques.

The measure of disclosure risk has been developed so much, whereas there has not been much work on the measure of data utility. For continuous data, measures of information loss are developed through discrepancies between some point estimates obtained original and protected data matrix (Domingo-Ferrer and Torra 2001; Yancey et al. 2002). Matrix discrepancy is measured by mean square error, mean absolute error and mean variation. For categorial data, Domingo-Ferror and Torra (2001) considered measures of information loss in three ways: direct comparison of categorial values, comparison of contingency tables and Entropy-based measures. Also some measures on distortion for contingency table are presented by Gomatam and Karr (2003), Shlomo and Young (2006) and Shlomo (2007).

Recently Woo et al. (2009) presented four global utility measures like propensity score measure, cluster analysis measure and two empirical CDF measures, which were constructed based on the distributions of the original and masked data. The empirical CDF measures proposed by Woo et al. (2009) are constructed on the basis of independent and identically distributed (IID) random variables. So, they

*Department of Statistics, University of Seoul, Seoul 130-743, South Korea

can be adaptable for the case of IID data, whereas they seem to be inappropriate for complex survey data because they are not independent and identical as well due to stratification, clustering and unequal selection probability.

In the paper, we focuses on complex survey data. Two data utility measures based on the weighted empirical distribution functions are proposed in section 2. Weighted empirical CDF is estimated from not only original and masked data but also survey weights attached to records. A simulation study conducted on 2006 Korea welfare panel data is presented in section 3. This study shows that the existing measures based on IID data are much lower than the proposed measures, which means that the existing measures may report data utility to be much more useful than actual usefulness in case of complex survey data. Finally, concluding remarks are given in section 4.

2. New CDF Utility Measures

2.1 Complex Survey Data

We consider a finite population with size N and each unit has p-dimensional values. A complex sample is selected through stratification, clustering and so on and some kind of adjustments including nonresponse adjustment and post-stratification adjustment are conducted. Let w_k be the final weight attached to the kth record. Then the final original microdata is as follows:

$$X = \{(w_k, x_{kj}) : k = 1, ..., m,\ j = 1, ..., p\} \tag{1}$$

where m is the number of records responded.

In order to release microdata to the public, we assume that statistical agencies use some sorts of statistical disclosure control techniques to the original microdata. The resulting masked microdata is represented as follows:

$$Z = \{(w_k, z_{kj}) : k = 1, ..., m,\ j = 1, ..., p\} \tag{2}$$

Here we assume that the original data x_{kj} are masked, but the weights w_k are not.

2.2 Existing CDF Measures for IID Data

If $x_k = (x_{k1}, x_{k2}, ..., x_{kp}),\ k = 1, ..., m$ are independent and identically distributed random data with distribution function $F(x)$, then the $F(x)$ can be estimated by the empirical distribution function $\hat{F}(x)$ such as

$$\hat{F}_X(t_1, ..., t_p) = \frac{1}{m} \sum_{k=1}^{m} I(x_{k1} \le t_1, ..., x_{kp} \le t_p) \tag{3}$$

where I is the indicator function such that $I(A) = 1$ if the condition A is satisfied and 0 if not.

Let u_k is the kth record of the combining original and masked data. Based on the above empirical CDF, Woo et al. (2009) suggested two data utility measures such as

$$DU_{max}(u) = max_{1 \le u_k \le 2m} |\hat{F}_X(u_k) - \hat{F}_Z(u_k)| \tag{4}$$

and

$$DU_{sq}(u) = \frac{1}{2m}\sum_{k=1}^{2m}[\hat{F}_X(u_k) - \hat{F}_Z(u_k)]^2 \tag{5}$$

Two measures above say that we lose much more information on masked data with larger values of DU_{max} and DU_{sq}.

2.3 New CDF Measures for Complex Survey Data

By the way, as we said earlier, our original microdata is not IID data, but complex survey data with different weights w_k. Our questions here are that (i) are two measures in (4) and (5) appropriate for complex survey data? (ii) If not, how do we make data utility measures like (4) and (5) appropriate for complex data? Our motivation is to use survey weights w_k in constructing data utility measures. We think that since survey weights play a vital role in survey estimation, the weights may play an important role in construction of utility measures. But until now its has not been considered in existing utility measures.

New data utility measures are made through the weighted empirical distribution function such as

$$\hat{F}_{XW}(t_1, ..., t_p) = \frac{\sum_{k=1}^{m} w_k I(x_{k1} \le t_1, ..., x_{kp} \le t_p)}{\sum_{k=1}^{m} w_k} \tag{6}$$

The weighted empirical CDF based on complex data having different survey weights is approximately unbiased for population CDF F. On the other hand, unweighted empirical CDF in (3) is biased for F. So we think two revised measures in the form of DU_{max} and DU_{sq} based on weighted empirical CDF may be more appropriate than the existing measures. Now our new utility measures through weighted empirical CDF in (6) are defined as follows:

$$DUW_{max}(u) = max_{1 \le u_k \le 2m}|\hat{F}_{XW}(u_k) - \hat{F}_{ZW}(u_k)| \tag{7}$$

and

$$DUW_{sq}(u) = \frac{1}{2m}\sum_{k=1}^{2m}[\hat{F}_{XW}(u_k) - \hat{F}_{ZW}(u_k)]^2 \tag{8}$$

where u_k is the kth record in the combining data.

Until now, we have not shown that these two measures are more appropriate than the counterparts theoretically. Instead we conducted a simulation study to say that empirically. The simulation study result is presented in the next section.

3. Simulation Study

In the simulation study, we used 2006 Korea welfare panel data surveyed by Korea Institute for Health and Social Affairs. Four variables were chosen among many survey variables such as x_1 is disposable income, x_2 cost, x_3 income, and x_4 tax per household were chosen. And after excluding missing values and outliers, we constructed a population with $m = 5,712$ records.

Three techniques are chosen as statistical disclosure control techniques. The first one is noise addition followed by Kim (1986). In this technique, multivariate normal noises are generated firstly as

Table 1: The average values of utility measures by sample sizes

SDC method	n	DU_{max}	DU_{sq}	DUW_{max}	DUW_{sq}
noise addition	100	0.1175	0.1995	0.0021	0.0068
	200	0.1044	0.1779	0.0015	0.0056
	300	0.0114	0.1703	0.0013	0.0053
rank swapping	100	0.0349	0.0812	0.0002	0.0004
	200	0.0268	0.0595	0.0001	0.0002
	300	0.0247	0.0523	0.0001	0.0002
microaggregation	100	0.0697	0.1299	0.0013	0.0018
	200	0.0593	0.1061	0.0011	0.0014
	300	0.0597	0.1037	0.0010	0.0014

$$\epsilon_k \ \sim \ MN_4(0, \alpha S), \ k = 1, ..., m$$

where $\epsilon_k = (\epsilon_{k1}, ..., \epsilon_{k4})$, S is the covariance matrix for 4 variables $x_k = (x_{k1}, ..., x_{k4})'$ and $\alpha(> 0)$ is a control parameter. Then protected data were made as follows

$$z_k = c(x_k - \epsilon_k) + (1 - c)\bar{x}$$

where $\bar{x} = (\bar{x}_1, ..., \bar{x}_4)'$ and $\bar{x}_j = \sum_{k=1}^{m} x_{kj}/m$ and $c = \sqrt{1/(\alpha + 1)}$.

As second and third techniques, rank swapping and microaggregation were chosen. For these techniques, we sorted each variable and made groups satisfying $|i/m - j/m| < \alpha$ where $0 < \alpha < 1$ and (i, j) are ranks of observations. Within a group, observations were swapped randomly in swapping technique and the average value was assigned to each observation in microaggregation technique.

From the population with size 5,712 we selected a sample with unequal probabilities with sample size m=100, 200 and 300 respectively. Next, for each selected data, three disclosure control techniques, noise addition, rank swapping and microaggregation were applied to the data and then obtained three set of masked data. Then four measures DU_{max}, DU_{sq}, DUW_{max} and DUW_{sq} in (4), (5), (7) and (8) were calculated. Finally, we repeated this process 1,000 times and obtained the average values of 1,000 values. The results are given in Table 1 and Table 2.

The simulation result says that as sample sizes increase then data utility increases. Next, as control parameter values *alpha* increase, then data utility decreases in all cases. Finally Between two measures, the value of new measures are greater than the value of existing counter part measures in all cases. Roughly speaking, the relationships are $DUW_{max} \approx 1.52 \times DU_{max}$ and $DUW_{sq} \approx 2.62 \times DU_{sq}$. Those mean that the existing measures report data utility to be more useful than actual usefulness in case of 2006 Korea welfare panel data.

4. Concluding Remarks

We considered complex data with unequal weights and new data utility measures based on weighted empirical CDF $\hat{F}_W$ were suggested. Bacause the weighted empirical CDF $\hat{F}_W$ is approximately unbiased estimator for the population CDF F, suggested measures based on $\hat{F}_W$ seem to be more appropriate than their counterpart measures in case of complex survey data. The simulation results showed that

Table 2: The average values of utility measures by parameter α

SDC method	α	DU_{max}	DU_{sq}	DUW_{max}	DUW_{sq}
noise addition	0.30	0.1043	0.1798	0.0014	0.0045
	0.50	0.1130	0.1941	0.0019	0.0066
	0.70	0.1194	0.2041	0.0023	0.0084
rank swapping	0.04	0.0200	0.0562	0.0001	0.0002
	0.06	0.0271	0.0654	0.0001	0.0003
	0.08	0.0330	0.0741	0.0001	0.0004
microaggregation	0.04	0.0448	0.0925	0.0004	0.0006
	0.06	0.0598	0.1152	0.0007	0.0010
	0.08	0.0745	0.1363	0.0012	0.0017

the values of DU_{max} and DU_{sq} are consistently lower than DUW_{max} and DUW_{sq} in case of 2006 Korea welfare panel data. Empirically speaking, we may over-report data utility for the complex survey data if we use existing data utility measures. So, one alternative is to use the proposed data utility measures, DUW_{max} and DUW_{sq} rather than DU_{max} and DU_{sq}. In addition, We need to investigate the appropriateness of DUW_{max} and DUW_{sq} over DU_{max} and DU_{sq} for the complex survey data theoretically. We remain that as a further study.

REFERENCES

Domingo-Ferrer, J., and Torra, V. (2001). "Disclosure protection methods and information loss for microdata," In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies,* eds. P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz : North-Holland, pp.91-110.

Gomatam, S. and Karr, A. (2003), "Distortion measures for categorical data swapping." Technical report number 131, National Institute of Statistical Sciences.

Shlomo, N. and Young, C. (2006). "Statistical disclosure limitation methods through a risk-utility framework," In *PSD' 2006 Privacy in Statistical Databases,* eds. J. Domingo-Ferror and L. Franconi, Springer LNCS 4302, pp. 68-81.

Shlomo, N. (2006). "Statistical disclosure limitation methods for contingency tables," *International Statistical Review,* 75, 199-217.

Woo M., Reiter, J.P. Oganian, A. and Karr (2009). "Global measures of data utility for microdata masked for disclosure limitation," *The Journal of Privacy and Confidentiality,* 1, 111-124.

Yancey, W.E., Winkler, W.E. and Creecy, R.H. (2002). "Disclosure risk assessment in perturbative microdata protection," Inference Control in Statistical Databases" eds. J. Domingo-Ferror, : Springer-Verlag, pp. 135-152.