CONSTRUCTION OF MASKING ERROR FOR CATEGORICAL VARIABLES

Gary Sullivan, Lilly Research Laboratories, and Wayne A. Fuller, Iowa State University
Wayne A. Fuller, Department of Statistics, ISU, Ames, IA 50011

ABSTRACT

A method of using random error to mask categorical variables is described. The procedure uses a random transformation of the original categorical variables into a normal vector. The procedure produces masked variables that have nearly the same percentage of items in each category as the original variables.

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

The expanding capacity of computers has produced an increasing demand for microdata. Government agencies such as the Census Bureau and the National Center for Health Statistics receive data requests from economic, business and medical researchers. These government agencies, and other data providers, are faced with the problem of supplying statistically useful data in a manner that minimizes the probability of revealing the identity of a respondent or confidential attributes of a respondent.

For example, suppose an agency releases tables of annual income cross-classified by sex and occupation for residents of Boone, Iowa. Furthermore, suppose Joseph Iostat is the only male statistician living in Boone, Iowa. Any user of the table, who knows that Joseph Iostat is the only male statistician in Boone, Iowa, can obtain the information about Joseph Iostat. Such an occurrence is called a case of attribute disclosure. Identity disclosure occurs if a user is able to link a respondent to the record of that respondent.

Data agencies must devise methods to reduce the possibilities of such disclosures in order to satisfy the pledges of confidentiality given to respondents. The methods employed to protect the anonymity of respondents are known as disclosure avoidance techniques.

1.2 Literature Review

Early references on confidentiality issues are Steinberg and Pritzker (1967) and Bachi and Baron (1969). Duncan and Lambert (1986) provide a good review of the federal statutes dealing with confidentiality. Mugge (1983) discusses confidentiality measures taken at the National Center for Health Statistics. Cox et al. (1985) provide a good discussion of Census Bureau data products and the techniques used to mask them before release.

Three forms of data release are frequency count tables, tables of aggregate magnitude data, and microdata files. Cell suppression, random data perturbation, random rounding and controlled rounding have been considered as possibilities to mask frequency count tables constructed from 1990 census data. Descriptions and examples of these techniques are presented in Cox et al. (1985). Also, see Fellegi (1975), Cox (1980), and Cox et al. (1986).

Sullivan and Fuller (1989) presented an algorithm for masking microdata. The algorithm masks continuous, discrete and classification variables. In this article we describe the masking procedure appropriate for classification variables. The objectives of the algorithm are:

1. To reduce the ability of an intruder to obtain attribute information about a particular respondent.
2. To maintain as nearly as possible the original structure of the data.

The formal manner in which we attempt to achieve these informal objectives will be discussed.

2. TRANSFORMATION OF CLASSIFICATION VARIABLES

The basic masking algorithm is designed for quantitative variables (continuous or discrete) and Bernoulli variables. To apply the algorithm to classification variables we first transform each classification variable into a set of Bernoulli variables.

Let $X(1), X(2), \ldots, X(n)$ be the responses to a variable, X, having categories $(C(1), C(2), \ldots, C(r))$. For each response, $X(t)$, $t=1, 2, \ldots, n$, we define a set of $r-1$ dummy variables $Z(t1), Z(t2), \ldots, Z(t,r-1)$ by

$$Z_{tj} = 1 \quad \text{if} \quad X_t = C_j ,$$
$$= 0 \quad \text{otherwise} . \tag{2.1}$$

Let $\phi(j)$ be the probability that an element is in category j, given that it is not in the first $j-1$ categories. Hence,

$$\phi_1 = \Pr\{C_1\} , \text{ and}$$
$$\phi_j = P_r\{C_j | \text{not } C_1, C_2, \ldots, C_{j-1}\} ,$$
$$\text{for } j=2, 3, \ldots, r-1 . \tag{2.2}$$

We further define the pseudo Bernoulli variables, $W(t1), W(t2), \ldots, W(t,r-1)$, by

$$W(t1) = Z(t1) ,$$
$$W(tj) = Z(tj) \quad \text{if } Z(ti) = 0 \text{ for } i<j \tag{2.3}$$
$$= 1 \quad \text{with prob. } \phi(j) \quad \text{if some } Z(ti) \neq 0 \text{ for } i<j ,$$
$$= 0 \text{ with prob. } [1 - \phi(j)] \quad \text{if some } Z(ti) \neq 0 \quad \text{for } i<j .$$

The $W(tj)$, $j=1, 2, \ldots, r-1$, are uncorrelated Bernoulli variables.

To mask the quantitative and Bernoulli variables, all observations are transformed to

standard normal variates using the sample univariate distribution functions. The quantitative variables and the Bernoulli variables are transformed by two slightly different algorithms. For a description of the transformation methods, see Sullivan (1989).

3. MASKING

3.1 Masking the Normal Data Vectors

The masked set of transformed data vectors are computed by adding a normally distributed error vector to each transformed vector of normal observations. Each error vector has mean zero and a covariance matrix approximately equal to a multiple of the covariance matrix of $\mathbf{Z}(t)$. Hence, the masked data are

$$\overset{*}{\mathbf{Z}}_t = \mathbf{Z}_t + \overset{*}{\mathbf{u}}_t \ , \quad t=1, 2, \ldots, n \tag{3.1}$$

where $\mathbf{u}^*(t)$ is a vector of random variables with mean $\mathbf{0}$ and covariance matrix approximately equal to a multiple of $\mathbf{m}(ZZ)$, where

$$\mathbf{m}_{ZZ} = n^{-1} \sum_{j=1}^{n} \mathbf{Z}_j' \mathbf{Z}_j \ . \tag{3.2}$$

Each $\mathbf{u}^*(t)$ is a function of $\mathbf{u}(t)$, a normal random vector having mean zero and near identity covariance matrix. For each $\mathbf{u}(t)$, we define $\mathbf{u}^*(t)$ as

$$\overset{*}{\mathbf{u}}_t = \alpha^{1/2} \mathbf{u}_t \mathbf{T}_{ZZ} \tag{3.3}$$

where $\mathbf{T}'(ZZ)\mathbf{T}(ZZ) = \mathbf{P}(ZZ)$, $\mathbf{P}(ZZ)$ is the correlation matrix of the transformed data, and $\alpha > 0$. The value of α is specified by the user. Then the initially masked data set is

$$\overset{*}{\mathbf{Z}}_t = \mathbf{Z}_t + \overset{*}{\mathbf{u}}_t = \mathbf{Z}_t + \alpha^{1/2} \mathbf{u}_t \mathbf{T}_{ZZ} \ , \quad t=1, 2, \ldots, n \ . \tag{3.4}$$

To insure that the transformed data vectors are adequately masked, we perform a distance check on the initially masked data. For each $\mathbf{Z}(t)$, let the statistical distance between $\mathbf{Z}(t)$ and $\mathbf{Z}^*(j)$ be denoted by $d(tj)$, for $j=1, 2, \ldots, n$. Let $d[tj(1)]$ be the smallest of the n distances and let $d[tj(2)]$ be the second smallest of the n distances. If the distance between $\mathbf{Z}(t)$ and $\mathbf{Z}^*(t)$, $d(tt)$, is too small relative to the other distances, $\mathbf{Z}^*(t)$ is declared to be inadequately masked. Then, a new $\mathbf{u}(t)$ is created with increased magnitude and a new $\mathbf{Z}^*(t)$ is recomputed. The distance check is again performed, and the process is repeated until the masked vector satisfies the distance requirement. The procedure only makes one remasking pass through the data. It would be possible to modify the program to make additional passes.

After the pass through the data to modify the error for the distance criterion, the masked Z-data are back transformed into masked X-data. Then an iterative procedure is used to improve the agreement between the correlation structure of the transformed variables and the correlation structure of the original variables. The error terms are adjusted in an attempt to achieve nearly identical correlations between the masked and original data. Details of this iterative process are provided in Sullivan (1989).

3.2. Back transforming the data to original scale

Let $\mathbf{Z}^* = [\mathbf{Z}^*(1), \mathbf{Z}^*(2), \ldots, \mathbf{Z}^*(p)]$ be the matrix of masked, transformed data, where

$$\overset{*}{\mathbf{Z}}_j = (\overset{*}{Z}_{1j}, \overset{*}{Z}_{2j}, \ldots, \overset{*}{Z}_{nj})' \tag{3.5}$$

is the vector containing the masked observations for the j-th transformed variable. To begin the back transformation, consider each vector of n observations, $\mathbf{Z}^*(j)$, separately. First, we define $\mathbf{R}^*(j)$ to be the vector of ranks of the n observations of $\mathbf{Z}^*(j)$, with the rank "n" being assigned to the largest value. We create a normalized version of the elements of $\mathbf{u}^*(j)$ by multiplying each $\mathbf{u}^*(tj)$ by

$$[(n-1)^{-1} \sum_{t=1}^{n} u^{*2}_{tj}]^{-1/2} \ .$$

The normalized vector form of $\mathbf{u}^*(j)$ is denoted by $\mathbf{u}^+(j)$. Further, we let $\mathbf{D}^*(j)$ be the vector of normalized and adjusted ranks of $\mathbf{Z}^*(j)$. The t-th element of $\mathbf{D}^*(j)$ is defined to be

$$\overset{*}{D}_{tj} = n^{-1}[\overset{*}{R}_{tj} + \psi(u^+_{tj})] \ , \tag{3.6}$$

for $t=1, 2, \ldots, n$, where $R^*(tj)$ is the rank of the t-th observation of the j-th variable and ψ is a continuous function mapping $u^+(tj)$ into $(-1, 0)$. The $\psi[u^+(tj)]$ values are perturbations to keep the $D^*(tj)$ values from being the simple ranks divided by n . After these computations are performed for all p variables, we have

$$\overset{*}{\mathbf{D}} = (\overset{*}{\mathbf{D}}_1, \overset{*}{\mathbf{D}}_2, \ldots, \overset{*}{\mathbf{D}}_p) \ , \tag{3.7}$$

where $\mathbf{D}^*(j) = [D^*(1j), D^*(2j), \ldots, D^*(nj)]'$ and $D(tj) \in (0, 1)$ for $t=1, 2, \ldots, n$, and $j=1, 2, \ldots, p$.

To convert the $D^*(tj)$ values to $X^*(tj)$ in the original scale, let $P(0k)$ be the mean of the k-th original Bernoulli variable. The masked Bernoulli value for $X(tk)$, $t=1, 2, \ldots, n$, is

$$\begin{aligned} \overset{*}{X}_{tk} &= 0 \quad \text{if} \quad \overset{*}{D}_{tk} \in (0, 1 - P_{0k}) \\ &= 1 \quad \text{if} \quad \overset{*}{D}_{tk} \in (1 - P_{0k}, 1) \end{aligned} \tag{3.8}$$

3.3. Back Transforming Classification Variables

The only remaining computation is to convert the sets of Bernoulli variables, created for the purpose of masking the classification variables, back to their categorical values. To mask the classification variable X having r categories, the Bernoulli variables $W(t1), \ldots, W(t,r-1)$ are created for each response, $X(t)$ (see Section 2.1). These responses to the

Bernoulli variables are masked and denoted by $W^*(t1), W^*(t2), \ldots, W^*(t,r-1)$. To determine masked categorical values, $X^*(1), X^*(2), \ldots, X^*(n)$, we first define $Z^*(tj)$ as

$$\begin{aligned} Z^*_{tj} &= W^*_{tj} && \text{for } j=1 \\ &= [1 - \sum_{i=1}^{j-1} Z^*_{tj}]W^*_{tj} && \text{for } j=2, \ldots, r-1 . \end{aligned} \qquad (3.9)$$

Then $X^*(t)$ is defined as

$$X^*_t = C_j \quad \text{if} \quad Z^*_{tj} = 1 , \quad t=1, \ldots, n , \qquad (3.10)$$

where $C(j)$ is the j-th category of the variable X .

In the computer algorithm, additional iterations are carried out to improve the agreement between the structure of the masked data and the correlation structure of the original variables.

4. EXAMPLE

We use the masking algorithm to mask a computer generated data set. The data set included a Poisson variable with parameter $\lambda = 1.8$, a standard normal variable highly correlated with the Poisson variable, and a classification variable having categories (1, 2, 3, 4). Large values of the Poisson and normal variables were associated with the first category of the classification variable, and small values of the normal and Poisson variables were associated with the fourth category of the classification variable.

We will use the masked data set to study the correlation structure of the data sets within subgroups of the classification variable. The normal and Poisson variables are positively correlated with each other and nearly uncorrelated with the chi-square variable. However, the data were generated so that the correlation structure is very different in different subgroups defined by the classification variable. We will demonstrate that in such cases the correlation structures within categories of the classification variable are not retained through the mask. That is, the masking operation preserves global second moment properties, but higher order properties may be distorted.

The computer generated data set consisted of 300 observation vectors. Each vector contained a normal, a chi-square, a Poisson and a classification variable. The original data set has four variables, but the analysis vector has six variables, because the classification variable is transformed into three Bernoulli variables. In the masking operation, the variance of the error term, defined by the value of α in (3.3), was set to 0.3.

In discussing the results of the mask, the variables $X(1)$, $X(2)$, $X(3)$, and $X(4)$ correspond to the standard normal, chi-square, Poisson and classification variables, respectively. We refer to the masked analogues of these variables as $X^*(1)$, $X^*(2)$, $X^*(3)$, and $X^*(4)$.

4.1. Cross tabulation of the classification variable

We begin by considering Table 4.1 which contains cross tabulations for the original and masked classification variables, $X(4)$ and $X^*(4)$. In masking the classification variables, we did not guarantee that the frequencies for the categorical values of the original variable are the same as those of the masked variable. However, we see from the table that the

Table 4.1. Frequencies for the original and masked classification

	X^*_4				
X_4	1	2	3	4	Total
1	64	4	4	9	81
2	4	19	3	4	30
3	5	3	37	8	53
4	8	7	6	115	136
Total	81	33	50	136	300

corresponding marginal row and column proportions are very similar. We also note that, of the 300 categorical values that were masked, 65 or 21.67% switched categories in the mask. The algorithm created a new set of responses which differs substantially from the original set.

4.2. Examination of overall correlation structure and cross correlations

We continue our analysis of the masked data set by examining the correlations of the quantitative variables, $X(1)$, $X(2)$, and $X(3)$. Let $\mathbf{R}(XX)$ be the sample correlation matrix of the original quantitative variables, $\mathbf{R}(X^*X^*)$ be the sample correlation matrix of the masked variables and $\mathbf{R}(XX^*)$ be the sample cross correlation matrix between the original and masked quantitative variables. The correlation matrices of the original and masked data sets are

$$\mathbf{R}_{XX} = \begin{pmatrix} 1 & -0.0964 & 0.8137 \\ -0.0964 & 1 & 0.0462 \\ 0.8137 & 0.0462 & 1 \end{pmatrix}$$

$$\mathbf{R}^{**}_{XX} = \begin{pmatrix} 1 & -0.0895 & 0.7941 \\ -0.0895 & 1 & 0.0468 \\ 0.7941 & 0.0468 & 1 \end{pmatrix}$$

We see that the correlation structures of the original and masked data sets are nearly identical.

The sample correlations between corresponding variables in the original and masked data sets

are

$$(r_{X_1\overset{*}{X}_1}, r_{X_2\overset{*}{X}_2}, r_{X_3\overset{*}{X}_3}) = (0.750, 0.758, 0.722)$$

The correlations are close to the target correlation of 0.74.

4.3. Comparisons within subgroups of the classification variable

We investigate the effect of masking the classification variable by looking at the structure of the quantitative variables within subgroups of the classification variable.

We begin by considering the means and standard deviations of the normal, chi-square and Poisson variables within the four subgroups of the classification variable. The subgroup means and standard deviations for both the original and masked data sets are found in Table 4.2. The frequency counts for each category, given in Table 4.3, are the number from which the descriptive statistics were computed.

Table 4.2. Category means and standard deviations for the normal, chi-square and Poisson variables

	Category			
Variable	1	2	3	4
X_1	0.764 (0.810)	0.261 (0.628)	0.369 (0.725)	-0.658 (0.848)
$\overset{*}{X}_1$	0.728 (0.833)	0.321 (0.776)	0.398 (0.718)	-0.661 (0.846)
X_2	1.232 (1.489)	0.450 (0.694)	0.652 (0.794)	1.148 (1.483)
$\overset{*}{X}_2$	1.163 (1.466)	0.646 (0.834)	0.659 (0.863)	1.168 (1.569)
X_3	2.975 (1.193)	1.933 (0.980)	2.358 (1.039)	0.882 (0.870)
$\overset{*}{X}_3$	2.827 (1.212)	2.242 (1.061)	2.340 (1.287)	0.912 (0.839)

Table 4.3. Frequency counts for categories of the original and masked classification variables

	Category			
Variable	1	2	3	4
X_4	81	30	53	136
$\overset{*}{X}_4$	81	33	50	136

As expected, we observe more disparity between corresponding variable means and standard deviations of the two data sets for subgroups having smaller frequency counts. For example, the subgroup means and standard deviations for the fourth category are very similar for the chi-square and Poisson variables, and nearly identical for the original and masked normal variables. The corresponding original and masked variable means and standard deviations for the second category differ much more than those of the fourth category. All differences are small relative to the standard errors.

Let us now investigate the correlation structure of the quantitative variables. Before giving the four correlation matrices corresponding to original data values within the subgroups of the classification variable, we explain why these matrices will differ.

The chi-square variable, which is the square of the normal variable, is nearly uncorrelated with the normal and Poisson variables. Within subgroups of the classification variable, however, the chi-square variable is correlated with the normal and Poisson variables. Within the first category, the chi-square variable is positively correlated with the other two variables. This follows from the fact that records belonging to the first category have large positive Poisson and normal values and, hence, large chi-square values. Records in the fourth category are associated with small values of the normal and Poisson variables. Hence, the chi-square variable is negatively correlated with the normal and Poisson variables in the category "4" subgroup of the classification variable. Though not interpreted easily, the correlation structures for the second and third categories also differ from the correlation structure of the entire data set. The four correlation matrices corresponding to data vectors belonging to the four categories of the classification variable are exhibited below, where $\mathbf{R}[XX(j)]$ denotes the correlation matrix of data vectors for which $X(4) = j$.

We anticipate that the correlation matrix for the j-th subgroup of the original data set will differ from the correlation matrix for the j-th subgroup of the masked data set for the following reason. In masking a data set, the algorithm adds error vectors to the transformed data vectors. The error vectors have a covariance matrix which is a multiple of the covariance matrix of the total data. That is, a transformed data vector which belongs to the first subgroup of the classification variable has an error vector added which has the same covariance as an error vector added to a transformed data vector belonging to the fourth subgroup. Hence, the correlation matrix for the original data in the j-th subgroup will differ from the correlation matrix of the masked data belonging to the j-th subgroup because the original correlations within subgroups are not equal to the overall correlation.

We give in Table 4.4 the correlation matrices for the original data and for the masked data for the four categories.

In this discussion of subgroup correlation matrices, we have demonstrated that a user of a

Table 4.4. Correlation matrices of original and masked data by category

Cat.	Original			Masked		
1	1.00	0.88	0.70	1.00	0.31	0.84
	0.88	1.00	0.64	0.31	1.00	0.32
	0.70	0.64	1.00	0.84	0.32	1.00
2	1.00	-0.33	0.57	1.00	-0.01	0.65
	-0.33	1.00	-0.18	-0.01	1.00	0.37
	0.57	-0.18	1.00	0.65	0.37	1.00
3	1.00	0.74	0.70	1.00	0.09	0.97
	0.74	1.00	0.58	0.09	1.00	-0.05
	0.70	0.58	1.00	0.97	-0.05	1.00
4	1.00	-0.78	0.73	1.00	-0.33	0.55
	-0.78	1.00	-0.45	-0.33	1.00	-0.04
	0.73	-0.45	1.00	0.55	-0.04	1.00

microdata release masked by our algorithm cannot be guaranteed that a non-random subset of the masked data will be statistically representative of the same non-random subset of the original data. Specifically, we focused on subsets defined by categories of a classification variable in the data set. We also saw that if the frequency count of the subgroup is relatively large, the subgroup mean and standard deviation of the masked data set tend to be similar to the original subgroup mean and standard deviation. However, we cannot expect the same agreement from the subgroup correlation matrices of the original and masked data sets. In general, when a data set is masked by our algorithm, statistical relationships between variables within a non-random subset are not preserved. In such cases, the sophisticated user can recover the correct covariance matrix using measurement error techniques. See Fuller (1987). Our example was extreme in that we constructed it to have very different correlation structures in different subsets.

ACKNOWLEDGEMENT

This research was partly supported by Joint Statistical Agreement JSA 90-7 with the U.S. Bureau of the Census.

REFERENCES

Bachi, R., and Baron, R. 1969. Confidentiality problems related to data banks. Bulletin of the International Statistical Institute 43:225-241.

Cox, L. H. 1980. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association 75:337-385.

Cox, L. H., Fagan, J. T., Greenberg, B., and Hemmig, R. 1986. Research at the Census Bureau into disclosure avoidance techniques for tabular data. In Proceedings of the Section on Survey Research Methods of the American Statistical Association, 388-393.

Cox, L. H., Johnson, B., McDonald, S., Nelson, D., and Vazquez, V. 1985. Confidentiality issues at the Census Bureau. Paper presented at the First Annual Research Conference of the Bureau of the Census, Washington, D.C.

Duncan, G. T., and Lambert, D. 1986. Disclosure-limited data dissemination. Journal of the American Statistical Association 81:10-28.

Fellegi, I. P. 1975. Controlled random rounding. Survey Methodology 1:123-135.

Fuller, W. A. 1987. Mesurement Error Models. John wiley, New York.

Mugge, R. H. 1983. Issues in protecting confidentiality in national health statistics. In Proceedings of the Section on Survey Research Methods of the American Statistical Association, 592-594.

Steinberg, J., and Pritzker, L. 1967. Some experiences with and reflections on data linkage in the United States. Bulletin of the International Statistical Institute, 786-808.

Sullivan, G. 1989. The use of added error to avoid disclosure in microdata releases. Unpublished Ph.D. Dissertation, Iowa State University, Ames, Iowa

Sullivan, G. and Fuller, W. A. 1989. The use of measurement error to avoid disclosure. In Proceedings of the Section on Survey Research Methods of the American Statistical Association