

Catalogue no. 11-522-XIE

**Statistics Canada International
Symposium Series - Proceedings**

**Symposium 2006 :
Methodological Issues in
Measuring Population Health**



2006



**Statistics
Canada**

**Statistique
Canada**

Canada

Masking for Discrete Variables

Myron J. Katzoff and Jay J. Kim¹

Abstract

The use of discrete variables having known statistical distributions in the masking of data on discrete variables has been under study for some time. This paper presents a few results from our research on this topic. The consequences of sampling with and without replacement from finite populations are one principal interest. Estimates of first and second order moments which attenuate or adjust for the additional variation due to masking of known type are developed. The impact of masking of the original data on the correlation structure of concomitantly measured discrete variables is considered and the need for the further development of results for analyses of multivariate data is discussed.

KEY WORDS: masking; multinomial data; disclosure protection; randomized response.

1. Introduction

Data on individual subjects obtained through national sample surveys, often called microdata, are frequently requested for the purpose of secondary data analyses in which the sample persons are the conceptual units of observation. Even though the researchers who request these data may have no interest in determining the identity of the survey respondents, the data records can contain sufficient information to enable the identification of individuals. Therein lays the potential risk of violating a promise of confidentiality for survey respondents. Masking techniques have been specifically developed to reduce this risk.

Masking procedures include data swapping, microaggregation and transformations of the response variables. Chief among our interests are procedures which preserve correlations between variables or, more generally, would not obliterate statistical relationships and data features extant in the unmasked data. Clearly, for parameter estimation, techniques which enable removal of the effects of masking by simple adjustments or transformations of estimators would have high utility. For discrete data, by masking we shall mean a method of disclosure protection in which the original responses for a sample member are changed in accordance with pre-specified transformations and/or stochastic mechanisms.

In the paragraphs that follow, we confine ourselves to a discussion of a procedure that one might apply with categorical variables: perturbation of the categorical variable by the addition of another categorical variable of the same dimension, a “noise” variable. The reader might note that Warner (1971) may have been the first to identify a connection between this method of masking for multinomial variables and randomized response. There is a subtle and important point regarding differences in the applications of randomized response and masking. For masking, as the suppliers of microdata, we would know the unaltered (*i.e.*, true) sample values but randomized response denies both the data supplier and the data user knowledge of those values. For this reason, the application of the probability mechanism in randomized response techniques requires that it be distributed independently of the so-called true values but, for masking, we can allow the mechanism to depend on them. The earlier work by Kim and Flueck (1977, 1978) is also of interest in a study of the approach considered in this paper which collates a few of the results currently available for multinomial data masked with multinomial noise, clarifies some points concerning statistical

¹Both authors are from the National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782, USA. **Disclaimer** This paper represents the views of the authors and should not be interpreted as representing the views, policies or practices of the Centers for Disease Control and Prevention, National Center for Health Statistics.

inference with masked categorical data, presents a few new results and indicates where some theoretical voids still exist.

2. Multinomial Masking and A Connection with Randomized Response

Kim and Flueck (1977), denoted by KF in this section, have considered randomized response techniques for the case of three or more categories in multinomial sampling. We reprove their result for the easily generalized trinomial case, making modest changes in notation, in order to (1) conform to the presentation of the portion of the developments for dichotomous variables in Kim (1987) which involved sampling with replacement and (2) clarify the relationship between this work and that of Gouweleuw, et. al. (1998), hereafter referred to as GKWW (1998). In their additive model, KF define independent discrete variables C_j , representing the original sample or true values, and a_j , the noise variables, for $j = 1, 2, \dots, n$ (the sample size) both of which take the values 1, 2 and 3. We compute and report $(C_j + a_j)_{\text{mod } 3}$. For $k = 0, 1, 2$, setting $\pi_k = \Pr\{C_j = k\}$ and $p_k = \{a_j = k\}$ for $j = 1, 2, \dots, n$ when nonzero $p_k \neq \frac{1}{3}$, KF supply unbiased estimators $\hat{\pi}_k$ for the π_k and variances of those estimators using straightforward linear algebra.

The numbers 0, 1 and 2 can be properly associated with the trinomial outcomes $\bar{v}_1 = (1, 0, 0)^T$, $\bar{v}_2 = (0, 1, 0)^T$ and $\bar{v}_3 = (0, 0, 1)^T$, respectively. Following GKWW (1998), we denote the original three-dimensional data-vectors by $\bar{\xi}_j$ and the perturbed data by \bar{x}_j . For each of these vectors, for $j = 1, 2, \dots, n$, we have the three possible outcomes \bar{v}_1 , \bar{v}_2 and \bar{v}_3 ; and if we let \bar{a} be one of these, then

$$[\bar{x}_j = \bar{a}] = [\bar{x}_j = \bar{a} \text{ and } \bar{\xi}_j = \bar{v}_1] \cup [\bar{x}_j = \bar{a} \text{ and } \bar{\xi}_j = \bar{v}_2] \cup [\bar{x}_j = \bar{a} \text{ and } \bar{\xi}_j = \bar{v}_3].$$

Form

$$\bar{T}_x = \sum_{j=1}^n \bar{x}_j = \left(\#[\bar{x}_j = \bar{v}_1], \#[\bar{x}_j = \bar{v}_2], \#[\bar{x}_j = \bar{v}_3] \right)^T,$$

where $\#[\bar{x}_j = \bar{v}_1]$ denotes the number of vectors \bar{x}_j in the sum for which $\bar{x}_j = \bar{v}_1$. We then have that

$$\bar{T}_x = \sum_{j=1}^n \begin{pmatrix} I_{[\bar{x}_j = \bar{v}_1 | \bar{\xi}_j = \bar{v}_1]} & I_{[\bar{x}_j = \bar{v}_1 | \bar{\xi}_j = \bar{v}_2]} & I_{[\bar{x}_j = \bar{v}_1 | \bar{\xi}_j = \bar{v}_3]} \\ I_{[\bar{x}_j = \bar{v}_2 | \bar{\xi}_j = \bar{v}_1]} & I_{[\bar{x}_j = \bar{v}_2 | \bar{\xi}_j = \bar{v}_2]} & I_{[\bar{x}_j = \bar{v}_2 | \bar{\xi}_j = \bar{v}_3]} \\ I_{[\bar{x}_j = \bar{v}_3 | \bar{\xi}_j = \bar{v}_1]} & I_{[\bar{x}_j = \bar{v}_3 | \bar{\xi}_j = \bar{v}_2]} & I_{[\bar{x}_j = \bar{v}_3 | \bar{\xi}_j = \bar{v}_3]} \end{pmatrix} \begin{pmatrix} I_{[\bar{\xi}_j = \bar{v}_1]} \\ I_{[\bar{\xi}_j = \bar{v}_2]} \\ I_{[\bar{\xi}_j = \bar{v}_3]} \end{pmatrix} \quad (1)$$

where $I_{[\bar{\xi}_j = \bar{a}]}$ is the indicator variable for the event $[\bar{\xi}_j = \bar{a}]$ and $I_{[\bar{x}_j = \bar{b} | \bar{\xi}_j = \bar{a}]}$ is the indicator variable for the event $[\bar{x}_j = \bar{b}]$ given that the event $[\bar{\xi}_j = \bar{a}]$ has occurred. It then follows that

$$E \left[\bar{T}_x \mid \{\bar{\xi}_j\}_{j=1}^n \right] = P^T \sum_{j=1}^n \begin{pmatrix} I_{[\bar{\xi}_j = \bar{v}_1]} \\ I_{[\bar{\xi}_j = \bar{v}_2]} \\ I_{[\bar{\xi}_j = \bar{v}_3]} \end{pmatrix}, \text{ where } P = \begin{pmatrix} p_0 & p_1 & p_2 \\ p_2 & p_0 & p_1 \\ p_1 & p_2 & p_0 \end{pmatrix}.$$

This yields the unbiased estimator of \bar{T}_ξ , $\hat{T}_\xi = (P^T)^{-1} \bar{T}_x$, derived by GKWW (1998) where $\bar{T}_x = \left(\#[\bar{\xi}_j = \bar{v}_1], \#[\bar{\xi}_j = \bar{v}_2], \#[\bar{\xi}_j = \bar{v}_3] \right)^T$. Unbiased estimates of the π_k are then obtained by dividing the coordinates of \hat{T}_ξ by n .

To obtain the unconditional expressions for variances and covariances, note that

$$\text{Var} \left(\hat{T}_x \mid \{ \bar{\xi}_j \}_{j=1}^n \right) = \sum_{k=1}^3 T_\xi(k) B_k,$$

where, for $k=1,2,3$, $T_\xi(k) = \#[\bar{\xi}_j = \bar{v}_k]$, $B_k = \text{Diag}(\bar{p}_k) - \bar{p}_k \bar{p}_k^T$ and \bar{p}_k^T is row k of P . Since $E[T_\xi(k)] = n p_{k-1}$ for $k=1,2,3$, it can be seen that

$$E \left\{ \text{Var} \left[\hat{T}_\xi \mid \{ \bar{\xi}_j \}_{j=1}^n \right] \right\} = (P^T)^{-1} \left(n \sum_{k=1}^3 p_{k-1} B_k \right) P^{-1} \quad (2)$$

where

$$P^{-1} = \Delta^{-1} \begin{pmatrix} p_0^2 - p_1 p_2 & p_2^2 - p_0 p_1 & p_1^2 - p_0 p_2 \\ p_1^2 - p_0 p_2 & p_0^2 - p_1 p_2 & p_2^2 - p_0 p_1 \\ p_2^2 - p_0 p_1 & p_1^2 - p_0 p_2 & p_0^2 - p_1 p_2 \end{pmatrix}$$

and $\Delta = p_0^3 + p_1^3 + p_2^3 - 3p_0 p_1 p_2$. Since $E \left[\hat{T}_\xi \mid \{ \bar{\xi}_j \}_{j=1}^n \right] = \bar{T}_\xi$,

$$\text{Var} \left\{ E \left[\hat{T}_\xi \mid \{ \bar{\xi}_j \}_{j=1}^n \right] \right\} = \text{Var} \left[\bar{T}_\xi \right] = n V(\bar{\pi}) \quad (3)$$

where $\bar{\pi} = (\pi_0, \pi_1, \pi_2)^T$ and $V(\bar{\pi}) = \text{Diag}(\bar{\pi}) - \bar{\pi} \bar{\pi}^T$. For the $\hat{\pi}_k$, the unconditional or total variance is n^{-2} times the sum of the matrices on the right-hand sides of equations (2) and (3).

The situation is much less complicated for the two-dimensional case in GKWW (1998), where

$P = \begin{pmatrix} \theta_0 & 1 - \theta_0 \\ 1 - \theta_1 & \theta_1 \end{pmatrix}$. It may be seen that

$$\text{Var} \left(\hat{T}_\xi \right) = \left(n\pi(1-\pi) + \frac{1}{\Delta^2} [n\pi\theta_0(1-\theta_0) + n(1-\pi)\theta_1(1-\theta_1)] \right) \bullet \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \text{ where } \Delta = \theta_0 + \theta_1 - 1.$$

If $\theta_0 = \theta_1 = p$, say, this expression is in complete agreement with Kim (1987) and Warner (1971).

3. Masking Conditional on Response

The work of Kim and Flueck (1978) on randomized response suggests that, for masking, we could consider four sampling combinations for the selection of sample cases and masking-variable values corresponding to selections made in accordance with simple random sampling with replacement (SRS WR) and simple random sampling without replacement (SRS WOR). We discuss only one of these combinations here, obtaining new results for sampling from a finite population: selection of sample cases using SRS WOR but masking employing SRS WR, which we call Bernoulli masking, for the two-dimensional case. For this combination, the only necessary modification of the

multinomial sampling variance results is multiplication of $Var\left\{E\left[\hat{T}_{\xi} \mid \{\bar{\xi}_j\}_{j=1}^n\right]\right\}$ by the factor $\left(\frac{N-n}{N-1}\right)$. Thus, for

the two-dimensional case, the term $n\pi(1-\pi)$ in $Var\left(\hat{T}_{\xi}\right)$ would be multiplied by this factor. Estimators for the first-order moments were developed in the last section. A new unbiased estimator of total variance is given below.

In a recent review of the Post Randomization Method (PRAM) procedure of GKWW (1998), Kim remarked that assessments of various masking procedures should be made using total variance and that inferences from masked

sample data require estimators of those variances. For the two-dimensional case, let $Z = \sum_{i=1}^n z_i$, where z_i is the masked value for sample member $i = 1, 2, \dots, n$ and

$$z_i = \begin{cases} 1, & \text{if member } i \text{ has the characteristic of interest} \\ & \text{(i.e., belongs to category 0)} \\ 0, & \text{otherwise} \end{cases}$$

Then an unbiased estimator for π , the proportion of population members with the characteristic is

$$\hat{\pi} = \frac{Z/n - (1 - \theta_1)}{(\theta_0 + \theta_1 - 1)}$$

and

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} \left(\frac{N-n}{N-1}\right) + \frac{\pi\theta_0(1-\theta_0) + (1-\pi)\theta_1(1-\theta_1)}{n(\theta_0 + \theta_1 - 1)^2}$$

An unbiased estimator of this variance is

$$\hat{V}(\hat{\pi}) = \frac{1}{nN(N-1)(\theta_0 + \theta_1 - 1)^2} \left[\frac{Z(n-Z)(N-n)}{n(n-1)} + (N^2 - 2N + n)[\hat{\pi}\theta_0(1-\theta_0) + (1-\hat{\pi})\theta_1(1-\theta_1)] \right]$$

Now consider the situation when sampling with arbitrary probabilities without replacement and when applying Bernoulli masking. Let γ_i denote the inclusion probabilities and γ_{ij} , the joint inclusion probabilities. If we redefine

Z by $Z = \sum_{i=1}^n z_i / \gamma_i$, then

$$\hat{\pi}^* = \frac{Z/N - (1 - \theta_1)}{\theta_0 + \theta_1 - 1}$$

is an unbiased estimator for π . An expression for the variance of $\hat{\pi}^*$ can be obtained by manipulating each of these two terms:

$$E\left\{Var\left[Z \mid \{\bar{\xi}_j\}_{j=1}^n\right]\right\} = \left(\sum_{i=1}^N \frac{\xi_i}{\gamma_i}\right)\theta_0(1-\theta_0) + \left(\sum_{i=1}^N \frac{(1-\xi_i)}{\gamma_i}\right)\theta_1(1-\theta_1)$$

and

$$\text{Var}\left\{E\left[Z \mid \left\{\xi_j\right\}_{j=1}^n\right]\right\} = \sum_{i=1}^N \sum_{j>i}^N (\gamma_i \gamma_j - \gamma_{ij}) \left(\frac{u_i}{\gamma_i} - \frac{u_j}{\gamma_j} \right)^2$$

where, as it is used here, ξ_i is the indicator variable for individual i defined by

$$\xi_i = \begin{cases} 1, & \text{if member } i \text{ has the characteristic of interest} \\ & \text{(i.e., belongs to category 0)} \\ 0, & \text{otherwise} \end{cases}$$

and $u_i = (\theta_0 + \theta_1 - 1)\xi_i + (1 - \theta_1)$, for $i = 1, 2, \dots, N$. Under SRS WOR, when $\gamma_i = \frac{n}{N}$ and $\gamma_{ij} = \frac{n(n-1)}{N(N-1)}$, these expressions simplify, as one would expect, to yield results that are consistent with those obtained earlier. It is more difficult to obtain an expression for an unbiased estimator of total variance for this situation than for the case of SRS WOR.

4. Second Order Moments for Pairs of Categorical Variables under Masking

In this section, we restate an important result of Kim (1987) using a more general notational framework than he used. The reader should consult Kim (1987) for proofs.

Consider a contingency table for this situation where $\{\pi_{ij} \mid i = 1, 2, \dots, K_1; j = 1, 2, \dots, K_2\}$ denote the cell probabilities.

If $\pi_i = \sum_{j=1}^{K_2} \pi_{ij}$ and $\pi_j = \sum_{i=1}^{K_1} \pi_{ij}$, a data user may have an interest in $\text{cov}(\hat{\pi}_i, \hat{\pi}_j)$ where the ‘‘hatted’’ quantities are estimators of marginal probabilities. For simple random sampling with replacement (SRS WR)

$$\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = \frac{\pi_{ij} - \pi_i \pi_j}{n}$$

but for simple random sampling without replacement (SRS WOR)

$$\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = \frac{\pi_{ij} - \pi_i \pi_j}{n} \left(\frac{N-n}{N-1} \right).$$

Now suppose that we introduce masking variables $Y_k, k = 1, 2, \dots, n$, selected under SRS WOR from a finite population of size $M \geq n$ where pM of the members equal one, $p \neq \frac{1}{2}$ and quantify

$$\begin{aligned} Z_{ik} &= (X_{ik} + Y_k)_{\text{mod } 2} \\ Z_{jk} &= (X_{jk} + Y_k)_{\text{mod } 2} \end{aligned}$$

where for $k = 1, 2, \dots, n$

$$X_{ik} = \begin{cases} 1, & \text{if observation } k \text{ belongs to cell } (i,j) \text{ for} \\ & \text{some } j=1, 2, \dots, K_2 \\ 0, & \text{otherwise} \end{cases} \quad X_{jk} = \begin{cases} 1, & \text{if observation } k \text{ belongs to cell } (i,j) \text{ for} \\ & \text{some } i=1, 2, \dots, K_1 \\ 0, & \text{otherwise} \end{cases}$$

Estimators of π_i and π_j are

$$\hat{\pi}_i = \frac{1}{n} \sum_{k=1}^n X_{ik} \quad \text{and} \quad \hat{\pi}_j = \frac{1}{n} \sum_{k=1}^n X_{jk}$$

and $\{(X_{ik}, X_{jk}) : k = 1, 2, \dots, n\}$ is the collection of sample values for cell (i,j) . If

$$F = \frac{N-n}{N-1} - 4p(1-p) \frac{N(M-n) - n(M-1)}{(N-1)(M-1)},$$

Kim (1987) has shown that under the above sampling and masking procedures,

$$\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = \frac{\pi_{ij} - \pi_i \pi_j}{n(1-2p)^2} \cdot F + \frac{p(1-p)}{n(1-2p)^2} \left(\frac{M-n}{M-1} \right) (1 - 2\pi_i - 2\pi_j + 4\pi_{ij}) \quad \text{and}$$

$$\text{Var}(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n(1-2p)^2} \cdot F + \frac{p(1-p)}{n(1-2p)^2} \left(\frac{M-n}{M-1} \right)$$

with a similar expression for $\text{Var}(\hat{\pi}_j)$. If $M = n$, these expressions simplify in an obvious way and it becomes clear that this masking scheme will preserve the correlation structure of the unmasked variables. If this property is important to the user, then a masking scheme like the one just described may be preferable to swapping or microaggregation. A more complete discussion of the topic covered in this section is given in Kim (1987). Depending upon the circumstances to be considered, even within the narrow range we have examined, some masking procedures seem to be better than others.

5. Future Research

Future research on masking for discrete variables should give consideration to the potential benefits of selecting WOR masking values from finite populations as in section 4. Further effort might also include studies of what can be done with multistage cluster designs and ways to reduce or eliminate the complexities of estimation that are encountered when sampling with unequal probabilities WOR. The feature of masking that allows the data-supplier the choice of making the randomization mechanism dependent on observed data values may prove to be valuable in these situations.

Clearly, variance estimators appropriate for various sampling procedures and masking strategies are necessary for inferences based upon information supplied to the users of masked data. However, in many situations, there may exist satisfactory approximations for those estimators for certain kinds of statistical inference and it would seem appropriate to study the properties of those estimators in such cases. These approximate estimators can be particularly important in situations where certain parameters used in masking must be estimated from the data in hand.

Finally, like so many others who have worked on techniques for disclosure protection, we recognize the need for measures of disclosure protection and, additionally, for criteria for determining what data elements should be masked or not.

References

- Kim, J. (1987), "A Further Development of the Randomized Response Technique for Masking Dichotomous Variables", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 239-244.
- Kim, J. and Flueck, J. (1977), "An Additive Randomized Response Model", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp.351-355.
- Kim, J. and Flueck, J. (1978), "Modification of the Randomized Response Techniques for Sampling without Replacement", *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp.346-350.
- Warner, S. (1971), "The Linear Randomized Response Model", *Journal of the American Statistical Association*, v.66, pp. 884-888.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf, P.-P. (1998), "Post Randomization for Statistical Disclosure Control: Theory and Implementation", *Journal of Official Statistics*, v.14, pp. 463-478. [Referred to in text by GKWW.]