A FURTHER DEVELOPMENT OF THE RANDOMIZED RESPONSE TECHNIQUE FOR MASKING DICHOTOMOUS VARIABLES

Jay Kim, Bureau of the Census*

Abstract

Survey data is sometimes released in microdata form. Due to the large amount of information on the microdata file, respondents are subject to reidentification risk. To reduce this risk, the microdata may be masked prior to its release. The randomized response technique, initially proposed as an interviewing instrument for collecting data on sensitive characteristics, was later suggested as a masking scheme. Among all the masking schemes for a discrete variable, only this scheme can preserve the correlation structure of the unmasked variables. This is a distinct advantage when multivariate statistical analysis is to be performed on the data. This paper deals with the methodology for protecting the variance, covariance and hence the correlation structure of the unmasked variables throughout the masking.

I. Introduction

Data gathered from a survey is sometimes released in the form of microdata. If someone has an access to an additional data file which has common information with the microdata file, respondents on the microdata file are subject to reidentification risk, and the confidentiality of the data is liable to be compromised. To reduce this risk, the microdata may be masked prior to its release.

A variety of methods have been proposed for masking. For masking discrete variables, data swapping [Dalenius, 1977], slicing [Paass and Wauschkuhn, 1985], subrecords combination [Paass and Wauschkuhn, 1985] and randomized response technique [Warner, 1965] are available.

All the above except for the randomized response technique (RRT) destroy the original correlation structure among the variables. This implies that the results of any statistical analysis which rely on the correlation structure performed on the masked data will be the same as those obtained from the unmasked data only if the masking is done using the RRT. In this regard, the RRT is the superior masking scheme. However, this technique has not been fully developed as a masking scheme for actual use. This paper is intended to fill this gap.

One of the early criticisms of the RRT was that correlations could not be estimated when at least one of two characteristics is estimated by RRT. This has prevented randomized response from being of practical value for use as a disclosure avoidance technique. This paper provides the methodology by which the covariance, thus correlation can be evaluated.

Warner (1971) is the first who proposed to use the RRT for masking discrete variables. Dalenius (1977) studied the RRT in more detail as a masking scheme. He proposed two different schemes in case of dichotomous variables. To illustrate how the RRT is employed, we present the following.

We define

$$x = \begin{cases} 1, & \text{if a respondent has a sensitive characteristic of interest,} \\ 0, & \text{otherwise} \end{cases}$$

Assuming a respondent selects 0 or 1 based on the probability basis to decide his/her response to the question, we define

$$y = \begin{cases} 1, & \text{if a respondent selects 1} \\ 0, & \text{otherwise} \end{cases}$$

Masking Scheme 1: Take $z = (x + y)_{\text{mod } 2}$
Define $\pi = Pr(x=1)$ and $p = Pr(y=1)$.
Then the estimator, $\hat{\pi}$, of π is

$$\hat{\pi} = \frac{\sum_{i=1}^{n} z_i/n - p}{1 - 2p}, \quad p \neq 1/2 \qquad (1)$$

Masking Scheme 2: Compare x and y, and set

$$z = \begin{cases} 1, & \text{if } x=y, \\ 0, & \text{otherwise.} \end{cases}$$

In this case

$$\hat{\pi} = \frac{\sum_{i=1}^{n} z_i/n - (1-p)}{2p - 1}, \quad p \neq 1/2 \ . \qquad (2)$$

This coding approach is equivalent to Warner's original randomized response design (1965).

Dalenius did not give in his paper the variance formula for the estimators in equations (1) and (2) maybe assuming that Warner's formula or a similar one can be used. The variance formula for both approaches are identical, hence if we assume the same settings as in Warner's scheme, we can use Warner's variance formula. However, a few points need to be addressed concerning the formula.

For illustration suppose that a bag of paper slips each bearing either 0 or 1 is used as a randomization device. Assume the probability of a slip bearing 1 is p. In Warner's scheme (or in coding approach), a respondent is directed to choose a slip and use the number on the slip to respond to the sensitive question and return the slip to the bag. If the proportion of the slips bearing 1 which were used for response, is calculated from a sample, it would be an estimate of p, rather than p itself. This implies that, in actuality, estimators in equations (1) and (2) are all ratio estimators, thus all biased. Since p can not be calculated from Warner's scheme, we have to resort to p. Thus, Warner's estimator and the estimator in equation (1) are at best approximate, and Warner's variance formula assuming $p = p$ is also approximate.

Only if the number of slips (denoted by M) in the bag is identical with the respondent sample size (n) and a used slip is not allowed to be returned to the bag (i.e., sampling without replacement), the exact p will be used.

Warner's scheme was developed as an interviewing instrument for the case of sensitive characteristics, but in the current situation, a scheme such as addition mod 2 is used to mask the data. Hence the size of M in the latter is more flexible than in the former. That is, if the number of slips is the same as the respondent sample size (i.e., M = n) and the slips are sampled without replacement, in Warner's interviewing situation the last interviewed person has no choice but the (remaining) last slip in the bag. Hence he/she may feel that the interviewer knows the number on the slip and hence his answer to the question can be decoded, leading to possible bias or nonresponse. This necessitates that the number of slips in the bag is greater than the respondent sample size. However, in the masking situation, maintaining M = n does not compromise the confidentiality of the file. Only if M = n and slips are sampled without replacement, the estimator proposed by Warner is exact and an exact variance and covariance formula for the estimator can be derived.

The following example shows an advantage of using the new formula. When two parameters of interest (π_1 and π_2) are .3 and .4, the joint probability of a respondent having both characteristics (π_{12}) is .1, p = .6, n = 1,000 and N = 10,000, where n and N are sample and population size, respectively, the correlation coefficient based on the proposed formula is -.0891 (according to equations (3) and (6)) which is exactly the same as the coefficient one can obtain from the unmasked data but the correlation based on the Warner's scheme is -.00321 (according to equations (4) and (7)) which is inexact and significantly different from that based on the unmasked data. In no case these two coefficients can be identical. Note that the correlation coefficient based on Warner's scheme is always inexact but the correlation based on the proposed approach is exact and also unbiased if the estimator before masking is unbiased.

II. Selection of Response

Following Warner's terminology, we will use "response" to represent the paper slip. All the slips in the bag constitute the "response population." To insure exactness of the estimator and variance estimator, throughout this article it is assumed that the response population is sampled without replacement and M = n.

In survey sampling, observations are usually made on a multitude of variables. Thus, it is most likely that more than one variable needs to be masked. If more than one dichotomous variable is masked, three different approaches can be taken for determining "response" for each respondent. In approach 1, generate a response population of size n, pick a slip for a respondent and use it for masking all the desired variables. In this approach a constant p is used for all variables. In the second approach, generate as many response populations as the number of variables to be masked and use a response population for masking a variable. According to this approach, if k dichotomous variables are to be masked, k responses (i.e., k 0 or 1 values), one response for each variable, need to be determined for each respondent. The third approach is a compromise between the first two approaches. In this approach, use a response population for more than one variable but not for all variables if the number of involved variables is at least 3.

The first approach is the simplest and least costly since it requires generation of only one response population and determination of 0 or 1 only once for each respondent disregarding the number of responses to be masked. One possible disadvantage of this approach is, if overall masking is not successful, a respondent can be identified. If an outside investigator has information on a variable for the identified respondent, he/she can find the added value. The investigator thus can decode every single masked value for the specific respondent. Therefore, the second approach is clearly the best in protecting data confidentiality and the third approach the second best.

III. Variance Estimation

As mentioned earlier, by sampling the response population without replacement and by using a new randomization device in which the number of responses in the device exactly matches the respondent sample size, an exact estimator and variance formula can be derived.

In the context of the variance formula, two populations - respondent and response populations need to be defined. The respondent population is the population from which a respondent sample is selected, while the response population is the population generated for masking the responses obtained from the respondent sample.

Let N denote the size of the respondent population, n the size of the respondent sample, and M the size of response population.

Usually, sampling is performed without replacement and a variance formula is derived for that situation. The formula for sampling with replacement is obtained by replacing the finite population correction factor in the former formula by 1. The variance estimator is obtained by substituting the estimator of π in the variance formula for the parameter.

Both schemes of Dalenius have the same variance formula even if the estimators are different. Here I will derive the variance formula based on scheme 1 as described above. In doing so, I will use the following algebraic identities:

$$z_i=(x_i + y_i)_{\text{mod } 2}=(1-x_i)y_i + x_i(1-y_i),\ i=1, 2,\ldots,n$$

The estimator, $\hat{\pi}$, of π is given in equation (1) and the variance of $\hat{\pi}$ is

$$V(\hat{\pi})= \frac{\pi(1-\pi)}{n(1-2p)^2}\left[\frac{N - n(1-2p)^2}{N - 1}\right],\ p\neq 1/2. \quad (3)$$

In the above the quantity in the brackets is the finite population correction factor. For the derivation of this formula, see Appendix 1. It is interesting to compare the above with Warner's formula:

$$V(\hat{\pi})_w = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(1-2p)^2}, \quad p \neq 1/2. \qquad (4)$$

Note that Warner derived this formula assuming both populations are sampled with replacement.

IV. Covariance Formula between Dichotomous Variables

In order to use the randomized response technique as a masking scheme, we need a formula for the covariance between the variables, at least one of which is masked. Given this formula, the users of the statistical software can correct for bias, if any, before inputting the estimates for some type of analysis. That is, most statistical packages can handle multiple regression or other analysis if a mean vector and either a variance-covariance matrix or a correlation matrix are input. Therefore, if the variance and/or covariance obtained from the masked data are biased and if a method for correcting for bias is known, then corrected variance-covariance matrix can be input for unbiased results.

Three situations need to be considered for the derivation of the formula: i) both variables are masked with the different p values; ii) both variables are masked with the same p values; and iii) only one variable is masked.

Case 1. Variables masked with different p-values

Denote p_1 and p_2 as the probabilities of a slip bearing 1 for variables 1 and 2, respectively, i.e., $E(y_{1i}) = p_1$ and $E(y_{2i}) = p_2$, i=1, 2,...,n.

Assume that the selection of a response from one response population is independent of the selection of a response from the other response population. Then

$$\mathrm{Cov}(\hat{\pi}_1,\hat{\pi}_2) = \frac{\pi_{12} - \pi_1\pi_2}{n}\left(\frac{N-n}{N-1}\right) \qquad (5)$$

where π_i, i=1, 2, is the probability of a respondent having characteristic i and π_{12} is the probability of a respondent having both characteristics.

Note that the above formula is identical with the formula in the usual direct sampling and unmasked-data situation. For the derivation of this formula, see Appendix 2.

Case 2. Both variables masked with the same p-value.

When both variables are masked by the same amount selected from the response population, the covariance formula is as follows;

$$\mathrm{Cov}(\hat{\pi}_1,\hat{\pi}_2) = \frac{\pi_{12} - \pi_1\pi_2}{n(1-2p)^2}\left[\frac{N - n(1-2p)^2}{N-1}\right], \quad p \neq 1/2. \qquad (6)$$

Note that this formula has the same finite population correction factor as the variance formula in equation (4). For the derivation of this formula, see Appendix 3. Compare the above with the covariance formula in Warner's situation:

$$\mathrm{Cov}(\hat{\pi}_1,\hat{\pi}_2) = \frac{\pi_{12} - \pi_1\pi_2}{n} + \frac{p(1-p)}{n(1-2p)^2}(1-2\pi_1-2\pi_2+4\pi_{12}). \qquad (7)$$

Note that this equation will be used for calculating correlation coefficients in Warner's case. This formula was developed not by Warner but by this author in this article. For derivation of this formula, see Appendix 3.

Case 3. When only one variable is masked

When the respondent population is sampled without replacement,

$$\mathrm{Cov}(\hat{\pi}_1,\hat{\pi}_2) = \frac{\pi_{12} - \pi_1\pi_2}{n}\left(\frac{N-n}{N-1}\right) \qquad (8)$$

This is the same covariance formula as for the unmasked variables. For the derivation of the formula, see Appendix 4.

V. Covariance Formula between a Dichotomous Variable and a Continuous Variable

Assume the first variable is a dichotomous variable masked by the randomized response technique and the second variable is a continuous variable. We will consider both the masked and unmasked continuous variable.

Case 1. Continuous variable is unmasked

The covariance formula obtained in the case 3 of section IV applies to this case. The only difference is π_2 in the formula is now the mean of the continuous variable, usually denoted by u_2 and π_{12} here is a total of the continuous variable for the respondents whose first variable has value 1.

Case 2. Continuous variable is masked

There are numerous ways of masking a continuous variable. In this section, we assume the variable is masked by the additive random noise approach (7). This approach can be defined as follows;

$$y_i = x_{2i} + e_i,$$

where x_{2i} is the second variable to be masked (for the i^{th} respondent) with $u = E(x_{2i})$ and $\sigma^2 = V(x_{2i})$, and e_i is the random noise added to x_{2i} (for the i^{th} respondent) which follows $N(0,c\sigma^2)$, c is a constant. When σ^2 is not known, e_i is generated such that $e_i \sim N(0,c\hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the

estimate of σ^2. Let $\bar{y} = \sum_i^n y_i / n$ then

$$\mathrm{Cov}(\hat{\pi}_1, \bar{y}) = \frac{\pi_{12} - \pi_1 u}{n} \left(\frac{N-n}{N-1}\right) + \frac{c\pi_{12}}{n} \left(\frac{N-n}{N-1}\right). \quad (9)$$

Note that the first term is the same as the one for case 1 and the second term is the addition due to masking. For the derivation of the formula, see Appendix 5.

VI. Correlation Coefficients between the Variables

When the correlation coefficient is estimated based on the variance and covariance formulas derived above, only the correlation estimated in case 2 of section IV will be unbiased. To show the difference between the correlations based on new method and those based on Warner's approach, correlations were calculated for eight sets of parameters which are shown below. The former were calculated assuming case 2 of section IV.

Table 1 Comparison of Warner's and Proposed Correlations -- n = 1,000, and same p values for all variables

π_1	π_2	π_{12}	p	Warner's	Proposed
.5	.3	.2	.6	.2006	.2182
.5	.3	.2	.7	.2026	.2182
.4	.3	.2	.6	.3984	.3563
.4	.3	.2	.7	.3935	.3563
.3	.5	.1	.6	-.2006	-.2182
.3	.5	.1	.7	-.2026	-.2182
.3	.4	.1	.6	-.0032	-.0891
.3	.4	.1	.7	-.0130	-.0891

In no case in the above the two correlations are the same. In some cases, the correlations are significantly different from each other. As mentioned previously, this set of proposed correlations is exact and unbiased but that of Warner's is only approximate thus biased. Thus, we can conclude that the new approach is better. The results of statistical analysis based on the biased estimates will be biased. As mentioned above, in cases other than case 2, the proposed correlations are also biased, but unlike the Warner's, the bias can be corrected either by dividing by or adding some constant. For example, when two dichotomous variables are masked by different p values, the estimator of the correlation coefficient is

$$\frac{\hat{\pi}_{12} - \hat{\pi}_1\hat{\pi}_2}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)\ \hat{\pi}_2\ (1-\hat{\pi}_2)}} \left(\frac{N-n}{N-1}\right) \times \left\{(N-1)\sqrt{\frac{(1-2p_1)^2(1-2p_2)^2}{[N-n(1-2p_1)^2][N-n(1-2p_2)^2)}}\right\}. \quad (10)$$

In the above, an unbiased estimator can be obtained by dividing the formula by the quantity in the bracelets.

VII. Concluding Remarks

In this article, the theory required for using the randomized response technique as a masking scheme has been developed. By correcting for bias, if any, unbiased results can be obtained from the masked data if full sample data are used. If subdomain statistics are required, masking can be performed over all mutually exclusive and exhaustive subdomains using the same p value. In this way, the subdomain statistics can be "preserved" just like in the full sample. The full sample statistics obtained from the data will still be unbiased or can be made unbiased.

Thus far, we have dealt with dichotomous variables only, but a method needs to be developed for finding the variance and covariance formula involving multichotomous variables.

References

Cochran, W.G. Sampling Techniques, 3rd ed., John Wiley and Sons, 1977.

Dalenius, T. (1977), Privacy Transformations for Statistical Information Systems, Journal of Statistical Planning and Inference 1, 73-86.

Dalenius, T. and Reiss, S.P., Data - Swapping - A Technique for Disclosure Control.

Horvitz, D.G., Greenberg, B.G. and Abernathy, J.R. (1976), Recent Developments in Randomized Response Designs. In A Survey of Statistical Design and Linear Models. J.N. Srivastava (ed.), 271-285.

Kim, J. and Flueck, J.A. (1978), Modification of the Randomized Response Techniques for Sampling without Replacement, Proceedings of the Section on Survey Research Methods, American Statistical Association, 346-350.

Paass, G. and Wauschkuhn, U. (1985), Datenzugang, Datenschutz and Anonymisierung, Oldensbourg Verlag (Munchen).

Spruill, N.L. (1983), Confidentiality and Analytic Usefulness of Masked Business Microdata, The Public Research Institute, Alexandria, Va.

Warner, S., (1965), Randomized Response, A Survey Technique for Eliminating Evasive Answer Bias, Journal of the American Statistical Association, 60, 63-69.

Warner, S. (1971), The Linear Randomized Response Model, Journal of the American Statistical Association, 66, 884-888.

Appendix 1. Derivation of $V(\hat{\pi})$

Since p is a constant in (1)

$$V(\hat{\pi}) = \frac{V(\Sigma z_i / n)}{(1 - 2p)^2}, \quad p \neq 1/2 .$$

Assuming the identical distribution of z_i, i=1,2,...,n, and noting that $z_i^2 = z_i$, the above reduces to

$$\frac{1}{n(1-2p)^2}\{E(z_i) + (n-1)\,E(z_i z_j) \underset{i \neq j}{} - n[E(z_i)]^2\}. \qquad (11)$$

In the above

$$E(z_i) = \pi + p - 2p\pi$$

$$E(z_i z_j) = E(x_i x_j)[4E(y_i y_j) - 4\,E(y_i) + 1] + E(y_i y_j)[1 - 4E(x_i)] + 2E(x_i y_j).$$

Using

$$E(x_i x_j) = \frac{\binom{N\pi}{2}}{\binom{N}{2}} = \frac{\pi(N\pi - 1)}{N-1}$$

and

$$E(y_i y_j) = \frac{\binom{Mp}{2}}{\binom{M}{2}} = \frac{p(Mp - 1)}{M-1},$$

equation (11) reduces to

$$\frac{\pi(1-\pi)}{n(1-2p)^2}\left[\frac{N-n}{N-1} - 4p\,\frac{N(M-n) - n(M-1)}{(N-1)(M-1)} + 4p^2\,\frac{N(M-n) - n(M-1)}{(N-1)(M-1)}\right] + \frac{p(1-p)}{n(1-2p)^2}\left(\frac{M-n}{M-1}\right).$$

Appendix 2. Derivation of $\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2)$ When Variables Are Masked Different p Values

Since p_1 and p_2 are constants,

$$\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2) = \frac{1}{n(1-2p_1)(1-2p_2)}[E(z_{1i}z_{2i}) + (n-1)E(z_{1i}z_{2j}) \underset{i \neq j}{} - n\,E(z_{1i})E(z_{2j})]. \qquad (12)$$

Since $E(x_{1i}x_{2i}) = \pi_{12}$, $E(x_{1i}x_{2j}) = \frac{N\pi_1\pi_2 - \pi_{12}}{N-1}$

and

$E(y_{1i}y_{2j}) = E(y_{1i}y_{2i}) = p_1p_2$, the above equation reduces to equation (5).

Appendix 3. Derivation of $\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2)$ When Variables Are Masked with Same p Value

From equation (12) and $p_1 = p_2 = p$,

$$\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2) = \frac{1}{n(1-2p)^2}[E(z_{1i}z_{2i}) + (n-1)E(z_{1i}z_{2j}) \underset{i \neq j}{} - n\,E(z_{1i})E(z_{2j})]. \qquad (13)$$

Different from Appendix 2,

$E(y_{1i}y_{2j}) = E(y_i^2) = p$. However, as before

$$E(x_{1i}x_{2j}) \underset{i \neq j}{} = \frac{N\pi_1\pi_2 - \pi_{12}}{N-1}$$

and $E(x_{1i}x_{2i}) = \pi_{12}$.

Plugging the above expressions in equation (13),

$$\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2) = \frac{\pi_{12} - \pi_1\pi_2}{n(1-2p)^2}\left[\frac{N-n}{N-1} - 4p\,\frac{N(M-n) - n(M-1)}{(N-1)(M-1)} + 4p^2\,\frac{N(M-n) - n(M-1)}{(N-1)(M-1)}\right] + \frac{p(1-p)}{n(1-2p)^2} \times \left(\frac{M-n}{M-1}\right)(1 - 2\pi_1 - 2\pi_2 + 4\pi_{12}).$$

By plugging M = n in the above, we can obtain equation (6). By using $N \to \infty$ and $M \to \infty$, we can obtain equation (7) from the above equation. Note that the sampling-with-replacement variance formula can be obtained by using $N \to \infty$ and $M \to \infty$ in the above.

Appendix 4. Derivation of $\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2)$ When Only One Variable Is Masked

Assuming only the first variable is masked,

$$\mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2) = \frac{1}{n^2(1-2p)}\mathrm{Cov}(\Sigma z_{1i}, \Sigma x_{2i}) = \frac{1}{n}[E(x_{1i}x_{2i}) + (n-1)E(x_{1i}x_{2j}) - n\pi_1\pi_2]. \qquad (14)$$

Using $E(x_{1i}x_{2i}) = \pi_{12}$

and

$$E(x_{1i}x_{2j}) \underset{i \neq j}{} = \frac{N\pi_1\pi_2 - \pi_{12}}{N-1},$$

equation (8) is obtained from equation (14).

Appendix 5. Derivation of $\mathrm{Cov}(\hat{\pi}_1, \bar{y})$

In this case,

$$\mathrm{Cov}(\hat{\pi}_1, \bar{y}) = \mathrm{Cov}(\hat{\pi}_1, \bar{x}_2) + \mathrm{Cov}(\hat{\pi}_1, \bar{e})$$

Now where $\bar{e} = \sum_i^n e_i/n$.

$$\mathrm{Cov}(\hat{\pi}_1, \bar{x}_2) = \frac{1}{n(1-2p)}[E(z_{1i}x_{2i}) + (n-1)E(z_{1i}x_{2j}) - n\,E(z_{1i})E(x_{2j})]. \quad (i \neq j)$$

From Appendix 4 and equation (8), the above reduces to the first term of equation (9).

Also

$$\mathrm{Cov}(\hat{\pi}_1, \bar{e}) = \frac{1}{n(1-2p)}[E(z_{1i}e_i) + (n-1)E(z_{1i}e_j) - n\,E(z_{1i})\,E(e_j)] \quad (i \neq j)$$

which reduces to the second term of equation (9).

*This paper reports the general results of research undertaken by Census Bureau Staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.