SUBPOPULATION ESTIMATION FOR THE MASKED DATA

Jay Kim, U.S. Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Microdata, Masking, Subgroup, Estimation

1. Introduction

Microdata is sometimes masked to protect confidentiality of the respondents in the file. The data can be masked as a whole by a masking scheme such as the additive noise approach. Using this type of masked data, mean, variance and covariance can be estimated for whole population. However, data users might be interested in the estimates for some subgroups. For example, if the data is of demographic nature of both male and female respondents over all different races, one might be interested in the income of the black female. However, at this moment no method is known for estimating mean, variance and covariance for subgroups from the masked data if the masking is done on the data set as a whole. Currently, if users want a masked data file from which he can obtain the subpopulation estimates, the data disseminating agency should mask each subgroup separately. This will make the masking operations time-consuming and costly. In this report, the author tries to correct this situation by developing formulae for the mean, variance and covariance for subgroups when data is masked by the additive noise approach or the additive noise plus transformation approach. As usual, these formulae will be developed in the context of population and the formulae for the sample estimators are derived from them.

To evaluate the performance of these estimates in real situations, a data set is masked as a whole and subpopulation estimates are calculated from the data set based on the formulae developed in this paper. These estimates are then compared with those obtained from the unmasked data. Findings from this empirical investigation are also reported.

II. Estimation Procedure for Subpopulation When Data Is Masked by Additive Noise Approach

Define

x is the variable of interest which is to be masked;
σ^2 is the variance of x;
e is the noise following a distribution with zero mean and variance $c\sigma^2$, where c is a constant;

and

y is the masked variable, i.e., $y = x + e$.

Consequently,

$$V(y) = (1 + c)\sigma^2$$
$$V(x) = V(y)/(1 + c)$$

and

$$V(e) = V(y) - V(x).$$

Now for a subgroup s, denote

x_s: the variable of interest;
σ_s^2: variance of x_s;
y_s: masked variable.

For example, when a researcher is interested in the earned income of White persons, x is the earned income, s is the subgroup White and y_s is the masked earned income for White persons. Since, when masking was performed, the noise was generated for the whole data file including the overall variances of the original unmasked variables, y_s can be expressed as follows:

$$y_s = x_s + e.$$

Thus

$$E(\bar{y}_s) = E(\bar{x}_s), \qquad (1)$$

and

$$V(y_s) = \sigma_s^2 + c\sigma^2.$$

Therefore

$$\sigma_s^2 = V(y_s) - c\sigma^2.$$

Using the fact that

$$V(y) = (1 + c)\sigma^2,$$

σ_s^2 can be simplified as

$$\sigma_s^2 = V(y_s) - \frac{c}{1 + c} V(y). \qquad (2)$$

This suggests that the variance of x in group s can be obtained by subtracting $\frac{c}{1 + c}$ times the overall variance of the masked variable from the variance of the masked variable calculated for the subgroup.

The covariance between two variables can also be obtained for the subpopulation when at least one of the two is masked.

For the derivation of the covariance formula, let

x_i be the unmasked variables of interest, $i = 1, 2$

e_i be the noise added to x_i such that $E(e_i) = 0$ and $V(e_i) = cV(x_i)$, $i = 1, 2,$

and

$y_i = x_i + e_i$, $i = 1, 2$, be the masked variable.

Let $\bar{x}_{is}$ and $\bar{y}_{is}$, $i = 1, 2$, be the means of the unmasked and masked variables for subgroup s, respectively.

Then $E(\bar{y}_{is}) = E(\bar{x}_{is})$, $i = 1, 2$.

Note that

$$\mathrm{Cov}(y_{1s}, y_{2s}) = E[(x_{1s}+e_1)(x_{2s}+e_2)] - E(x_{1s}+e_1)E(x_{2s}+e_2). \quad (3)$$

Since e_i, $i = 1, 2$, are generated such that they are independent of x_i, $i = 1, 2$, if e_1 and e_2 are independent, the right side of the equation (3) becomes

$$\mathrm{Cov}(x_{1s}, x_{2s}).$$

Thus when e_1 and e_2 are independent,

$$\mathrm{Cov}(x_{1s}, x_{2s}) = \mathrm{Cov}(y_{1s}, y_{2s}). \quad (4)$$

If e_1 and e_2 are correlated such that

$$\mathrm{Cov}(e_1, e_2) = c\mathrm{Cov}(x_1, x_2),$$

then the right side of equation (3) becomes

$$\mathrm{Cov}(y_{1s}, y_{2s}) = \mathrm{Cov}(x_{1s}, x_{2s}) + c\mathrm{Cov}(x_1, x_2).$$

Thus

$$\mathrm{Cov}(x_{1s}, x_{2s}) = \mathrm{Cov}(y_{1s}, y_{2s}) - c\mathrm{Cov}(x_1, x_2).$$

But

$$\mathrm{Cov}(x_1, x_2) = \frac{1}{1+c}\mathrm{Cov}(y_1, y_2).$$

Hence

$$\mathrm{Cov}(x_{1s}, x_{2s}) = \mathrm{Cov}(y_{1s}, y_{2s}) - \frac{c}{1+c}\mathrm{Cov}(y_1, y_2). \quad (5)$$

This covariance formula is essentially identical to the variance formula in equation (2). That is, we can obtain the covariance formula in the above equation by replacing the variance terms in equation (2) by the corresponding covariance terms. If only one variable, x_1, say, is masked, then

$$\mathrm{Cov}(y_{1s}, x_{2s}) = \mathrm{Cov}(x_{1s}, x_{2s}). \quad (6)$$

Thus, users can obtain the covariance between two unmasked variables for a subgroup s from the covariance between two variables, only one of which is masked, without any adjustment. This is different from the case in which both variables are masked as seen in equation (5).

The sample estimators of the mean, variance and covariance can be obtained by replacing the population parameters by sample estimators in equations (1) through (6).

Note that, in general, the equality of equations (1) through (6) would not hold when the population parameters are replaced by sample estimates. Thus, to empirically compare two sides of equations (1), (2), (5) and (6), i.e. to empirically investigate the performance of the mean, variance and covariance formulae for subpopulations, microdata was masked as a whole and means, variances and covariances were calculated for subgroups before and after masking. Since random noise is generated for more than one unmasked variable, i.e., one set of noise for each variable, random noise can be generated in a fashion such that the sets are either correlated or not. Thus, for this report, the covariances were estimated twice, first from the data set with correlated noise and second from the data set with uncorrelated noise. The data set used for this study is 1980 Decennial Census tract data. The characteristics considered for this study are the age of the householder (age), the percent of persons aged 65 and above in the tract where the household is located (oldage), the median age of the houses in the tract mentioned above (medage), and finally, the median rent of the rental units in the tract (medrnt). The subgroups considered here are non-Spanish White and Asian. For these groups, estimated mean and variance ratios of the masked to the unmasked data are given in Tables 1 and 2, corresponding estimated covariance ratios are shown in Tables 3 and 4 when correlated sets of noise were used and the same in Tables 5 and 6 when independent sets of noise were used.

As in Tables 1 and 2, the means of the masked data are extremely close to those of the unmasked data. The differences between the means are all less than 1% of the means of the unmasked data, disregarding the race groups. More specifically, the differences are less than .2% for all four variables in the case of non-Spanish White.

For non-Spanish Asian, the differences are slightly bigger, but they are all less than .5%. This shows that the means of the masked data are almost always virtually identical to the true means.

The performance of the variance formula is not as good as that of the mean formula. For non-Spanish White, the difference ratio, i.e., (masked variance - unmasked variance)/(unmasked variance) is less than 1% for three out of four items and the remaining one is 1.4%. In the case of non-Spanish Asian, the ratio is less than 1% for three variables, ranging from .1% to .8%, but the last variable has a difference ratio of 1.12%. This shows that the variance estimates are also excellent.

The performance of the covariance formula is not as good as that of the variance formula. As shown in Table 3, three out of six covariances with the correlated noise for non-Spanish White are less than or equal to 1% off, two are about 2% off and one is 4% off from the ones from the unmasked data. Table 4 shows that, in the case of non-Spanish Asian, two out of six covariances are less than 1% off and the rest are 1.1% to 4.1% off.

When the uncorrelated noise is used as shown in tables 5 and 6, for non-Spanish White, two covariance ratios are less than 1% off and the rest are 1.3 to 4.1% off, and for non-Spanish Asian, two are less than 1% off and the remaining ones are 1.7% to 4% off.

Comparing Table 3 with Table 5 and Table 4 with Table 6, we can observe that the covariances obtained from the data with the correlated noise are slightly better than the corresponding ones from the data with the uncorrelated noise.

III. Estimation Procedure for Subpopulation When Data is Masked by Additive Noise Plus Transformation Approach

This masking scheme requires transforming y such that

$$z = ay + (1-a)\bar{y}$$

where z is the new masked variable and $a \doteq \sqrt{1/(1+c)}$. For this model and the derivation of the approximate formula of "a", see the reference.

Let z_s be z for subgroup s.

Since masking is assumed to be done for whole group, z_s can be expressed as

$$z_s = ay_s + (1-a)\bar{y} = a(x_s+e) + (1-a)\bar{y}.$$

Thus

$$E(\bar{z_s}) = aE(\bar{x_s}) + (1-a)E(\bar{z})$$

which is since $\bar{z} = \bar{y}$.

Hence

$$E(\bar{x_s}) = [E(\bar{z_s}) - (1-a)E(\bar{z})]/a. \qquad (7)$$

Also,

$$V(z_s) = a^2[V(x_s+e)] + 2Cov[a(x_s+e), (1-a)(\bar{x}+\bar{e})] \qquad (8)$$

$$+ (1-a)^2[V(\bar{x}+\bar{e})].$$

Now since

$$Cov(x_s, \bar{x}) = V(x_s)/n \quad \text{and} \quad Cov(e, \bar{e}) = cV(x)/n$$

assuming $V(z) = V(x)$, the right side of equation (8) reduces to

$$V(z) = V(x_s)[a^2 + \frac{2a(1-a)}{n}] + V(z)[a^2c + \frac{2a(1-a)c}{n} + \frac{(1-a)^2(1+c)}{n}].$$

Thus, $V(x_s)$ can be expressed as

$$V(x_s) = \{V(z_s) - V(z)[a^2c + \frac{2a(1-a)c + (1-a)^2(1+c)}{n}]\} / [a^2 + \frac{2a(1-a)}{n}]. \qquad (9)$$

If n is large

$$V(x_s) \doteq \frac{V(z_s)}{a^2} - cV(z) \qquad (10)$$

The covariance formula between two masked variables can be found in the same fashion as for the additive noise without transformation.

Let

$$z_i = a(x_i + e_i) + (1-a)(\bar{x}_i + \bar{e}_i), \qquad i = 1, 2.$$

For subgroup s, we define

$$z_{is} = a(x_{is} + e_i) + (1-a)(\bar{x}_i + \bar{e}_i), \qquad i = 1, 2.$$

Then

$$Cov(z_{1s}, z_{2s}) = a^2Cov(x_{1s} + e_1, x_{2s} + e_2) \qquad (11)$$

$$+ (1-a)^2Cov(\bar{x}_1 + \bar{e}_1, \bar{x}_2 + \bar{e}_2)$$

$$+ 2a(1-a)Cov(x_{1s} + e_1, \bar{x}_2 + \bar{e}_2).$$

If e_1 and e_2 are independent,

$$Cov(x_{1s} + e_1, x_{2s} + e_2) = Cov(x_{1s}, x_{2s}),$$

$$Cov(\bar{x}_1 + \bar{e}_1, \bar{x}_2 + \bar{e}_2) = \frac{1}{n}Cov(x_1, x_2)$$

and

$$Cov(x_{1s} + e_1, \bar{x}_2 + \bar{e}_2) = \frac{1}{n}Cov(x_{1s}, x_{2s}).$$

Thus the right side of equation (11) becomes

$$[a^2 + \frac{2a(1-a)}{n}] Cov(x_{1s}, x_{2s})$$

$$+ \frac{(1-a)^2}{n} Cov(x_1, x_2). \qquad (12)$$

Now since

$$Cov(z_1, z_2) = [a^2 + \frac{2a(1-a) + (1-a)^2}{n}] Cov(x_1, x_2), \qquad (13)$$

From equation (13), $Cov(x_1, x_2)$ can be expressed as a function of $Cov(z_1, z_2)$. By substituting this new expression for $Cov(x_1, x_2)$ in equation (12) and rearranging the terms in equation (12), we obtain

$$Cov(x_{1s}, x_{2s}) = \{Cov(z_{1s}, x_{2s}) - \frac{(1-a)^2}{[na^2 + 2a(1-a)+(1-a)^2]}Cov(z_1, z_2)\} / [a^2 + \frac{2a(1-a)}{n}].$$

If n is large, the above reduces to

$$Cov(x_{1s}, x_{2s}) \doteq Cov(z_{1s}, z_{2s}) / a^2. \qquad (14)$$

If e_1 and e_2 are correlated such that

$$Cov(e_1, e_2) = cCov(x_1, x_2)$$

then from equation (11)

$$\mathrm{Cov}(z_{1s}, z_{2s}) =$$

$$a^2[\mathrm{Cov}(x_{1s}, x_{2s}) + c\mathrm{Cov}(x_1, x_2)] \qquad (15)$$

$$+ (1-a)^2[\frac{1}{n}\mathrm{Cov}(x_1, x_2) + \frac{1}{n} c\mathrm{Cov}(x_1, x_2)]$$

$$+ 2a(1-a)[\frac{1}{n}\mathrm{Cov}(x_{1s}, x_{2s}) + \frac{1}{n} c\mathrm{Cov}(x_1, x_2)]$$

which is since

$$\mathrm{Cov}(x_{1s}, x_2) = \frac{1}{n}\mathrm{Cov}(x_{1s}, x_{2s}).$$

By combining terms, equation (15) can be reexpressed as

$$\mathrm{Cov}(z_{1s}, z_{2s}) = [a^2 + \frac{2a(1-a)}{n}] \mathrm{Cov}(x_{1s}, x_{2s}) \qquad (16)$$

$$+ [a^2c + \frac{2a(1-a)c + (1-a)^2(1+c)}{n}] \mathrm{Cov}(x_1, x_2).$$

Since

$$\mathrm{Cov}(z_1, z_2) =$$

$$(1+c) \{a^2 + [2a(1-a) + (1-a)^2]/n\} \mathrm{Cov}(x_1, x_2)$$

$\mathrm{Cov}(x_{1s}, x_{2s})$ can be derived from equation (16) as

$$\mathrm{Cov}(x_{1s}, x_{2s}) = \{\mathrm{Cov}(z_{1s}, z_{2s}) -$$

$$\frac{na^2c + 2a(1-a)c + (1-a)^2(1+c)}{(1+c)[na^2 + 2a(1-a) + (1-a)^2]} \qquad (17)$$

$$\mathrm{x} \quad \mathrm{Cov}(z_1, z_2)\} / [a^2 + \frac{2a(1-a)}{n}].$$

If n is large, the above reduces to

$$\mathrm{Cov}(x_{1s}, x_{2s}) \doteq$$

$$[\mathrm{Cov}(z_{1s}, z_{2s}) - \frac{c}{1+c} \mathrm{Cov}(z_1, z_2)]/a^2. \qquad (18)$$

If the second variable is not masked,

$$\mathrm{Cov}(z_{1s}, x_{2s}) =$$

$$E\{[a(x_{1s} + e_1) + (1-a) (\overline{x}_1 + \overline{e}_1)]x_{2s}\} \qquad (19)$$

$$- E[a(x_{1s} + e_1) + (1-a) (\overline{x}_1 + \overline{e}_1)]E(x_{2s})$$

$$= a\mathrm{Cov}(x_{1s} + e_1, x_{2s}) + (1-a)\mathrm{Cov}(\overline{x}_1 + \overline{e}_1, x_{2s})$$

$$= a\mathrm{Cov}(x_{1s}, x_{2s}) + \frac{(1-a)}{n} \mathrm{Cov}(x_1, x_2) .$$

Now

$$\mathrm{Cov}(z_1, x_2) = (a + \frac{1-a}{n}) \mathrm{Cov}(x_1, x_2) . \qquad (20)$$

Thus, from equations (19) and (20)

$$\mathrm{Cov}(x_{1s}, x_{2s}) =$$

$$[\mathrm{Cov}(z_{1s}, x_{2s}) - \frac{1-a}{na + 1-a} \mathrm{Cov}(z_1, x_2)]/a. \qquad (21)$$

The sample estimators of the mean, variance and covariance can be obtained by replacing the population parameters with sample estimators in equations (7) through (21).

The performance of the formulae for the mean, variance and covariance for a subpopulation was also empirically investigated by masking the same data set as before based on the additional noise plus transformation approach. Note that the new data set was created by imposing a linear transformation on the data set masked by the additive noise approach. Thus all the properties of the latter including those for the subpopulation except for "density of data points" are transferred to the former, which becomes clear when we compare Table 1 with Table 7, Table 2 with Table 8, Table 3 with Table 9, Table 4 with Table 10, etc. The differences in ratios between the tables are all zero. Thus what was observed in Tables 1 to 6 for the data masked by the additive noise approach almost exactly applies to Tables 7 to 12 for the data masked by the additive noise plus transformation approach.

IV. Concluding Remarks

In the past, estimation methods were not available for subgroups when the data was masked as a whole, thus necessitating either masking each subgroup separately or the calculation of ratios by the data releasing agency of the variance of the masked data to that of the unmasked data for the subgroups of interest and releasing them along with the masked data file. However, with the formulae developed above, the estimates of the mean, variance and covariance can be calculated for any subgroup. Also when data is masked on a subgroup basis, the mean, variance and covariance can be estimated for subgroups of the subgroups, i.e., sub-subgroups.

In the above, the mean, variance and covariance formulae were developed for subpopulations and their estimators were derived when the data were masked as a whole by either the additive noise approach or the additive noise plus transformation approach. For both approaches, the estimates of the mean, variance and covariance were calculated from the masked data for two race groups, non-Spanish White and non-Spanish Asian, and compared with estimates from the unmasked data. Both approaches of masking rendered almost identical results. It also should be mentioned that both the correlated and uncorrelated noise approaches were tried on the data. The correlated noise approach provided somewhat better results. Thus for the purpose of summarizing the results of the empirical investigation, we will concentrate on the results of the additive noise approach with correlated noise for both race groups.

The estimates of the means were the best among the three types of estimates, followed by the estimates of the variances and finally by the estimates of the covariances. The estimates of the means were virtually identical to those from the unmasked data. The estimates of the variances were excellent, the maximum difference ratio being only 1% which occurred only twice in eight chances. The estimates of the covariances were good whose maximum difference ratio being 4% which occurred twice in twelve cases.

All in all, these findings lead to the conclusion that it is safe to mask the whole data set once and to let the users estimate the subpopulation parameters such as the mean, the variance and the covariance based on the formulae developed here.

Reference

Kim, Jay Jong-IK (1986): "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation," Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 370-374.

Appendix

Table 1
Mean and Variance Ratios of the Masked Data to Those of the Unmasked for Non-Spanish White - Additive Noise Approach
c=.25 and n=29,079

	Age	Oldage	Medage	Medrnt
Mean Ratio	1.000	1.001	1.002	1.001
Variance Ratio	1.002	.993	1.000	1.014

Table 2
Mean and Variance Ratios of the Masked Data to Those of the Unmasked for Non-Spanish Asian - Additive Noise Approach
c=.25 and n=6,614

	Age	Oldage	Medage	Medrnt
Mean Ratio	1.000	1.004	1.004	.997
Variance Ratio	.995	1.012	1.008	.999

Table 3
Covariance Ratios of the Data Masked with Correlated Noise to Those of the Unmasked Data for Non-Spanish White - Additive Noise Approach
c=.25 and n=29,079

	Oldage	Medage	Medrnt
Age	1.023	1.010	.996
Oldage		.984	.960
Medage			.992

Table 4
Covariance Ratios of the Data Masked with Correlated Noise to Those of the Unmasked Data for Non-Spanish Asian - Additive Noise Approach
c=.25 and n=6,614

	Oldage	Medage	Medrnt
Age	.976	1.004	.989
Oldage		1.041	1.003
Medage			1.025

Table 5
Covariance Ratios of the Data Masked with Uncorrelated Noise to Those of the Unmasked Data for Non-Spanish White - Additive Noise Approach

	Oldage	Medage	Medrnt
Age	1.024	1.004	.967
Oldage		.987	.959
Medage			.996

Table 6
Covariance Ratios of the Data Masked with Uncorrelated Noise to Those of the Unmasked Data for Non-Spanish Asian - Additive Noise Approach

	Oldage	Medage	Medrnt
Age	.991	1.006	.974
Oldage		1.040	.983
Medage			1.024

Table 7
Mean and Variance Ratios of the Masked Data to Those of the Unmasked Data for Non-Spanish White - Additive Noise Plus Transformation
c=.25 and n=29,079

	Age	Oldage	Medage	Medrnt
Mean Ratio	1.000	1.001	1.002	1.001
Variance Ratio	1.002	.993	1.000	1.014

Table 8
Mean and Variance Ratios of the Masked Data to Those of the Unmasked Data for Non-Spanish Asian - Additive Noise Plus Transformation
c=.25 and n=6,614

	Age	Oldage	Medage	Medrnt
Mean Ratio	1.000	1.004	1.004	.997
Variance Ratio	.995	1.012	1.008	.999

Table 9
Covariance Ratios of the Data Masked with Correlated Noise to Those of the Unmasked Data for Non-Spanish White - Additive Noise Plus Transformation Approach
c=.25 and n=6,614

	Oldage	Medage	Medrnt
Age	1.023	1.010	.996
Oldage		.984	.960
Medage			.992

Table 10
Covariance Ratios of the Data Masked with Correlated Noise to Those of the Unmasked Data for Non-Spanish Asian - Additive Noise Plus Transformation Approach
c=.25 and n=6,614

	Oldage	Medage	Medrnt
Age	.976	1.004	.989
Oldage		1.041	1.003
Medage			1.025

Table 11
Covariance Ratios of the Data Masked with Uncorrelated Noise to Those of the Unmasked Data for Non-Spanish White - Additive Noise Plus Transformation Approach
c=.25 and n=29,079

	Oldage	Medage	Medrnt
Age	1.024	1.004	.967
Oldage		.987	.959
Medage			.996

Table 12
Covariance Ratios of the Data Masked with Uncorrelated Noise to Those of the Unmasked Data for Non-Spanish Asian - Additive Noise Plus Transformation Approach
c=.25 and n=6,614

	Oldage	Medage	Medrnt
Age	.991	1.006	.974
Oldage		1.040	.983
Medage			1.024

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.