SOME STATISTICAL PROPERTIES OF MICROAGGREGATED ESTIMATES

Jay Kim, Bureau of the Census [1/]
Washington, D.C. 20233

KEY WORDS: microaggregation, equal group, unequal group, correlation.

I. Introduction

Survey data is sometimes released in the form of microdata. Since the amount of information on the file can be large, there is a potential for disclosure for the respondents on the file. Many methods for protecting those on the file have been suggested. One method which has been suggested for the release of microlevel information is the creation of microaggregation files (see Govoni-Waite, 1985 and Wolf, 1988). There are a variety of methods to create a microaggregation file. However, under each method, much of the basic statistical information is distorted from that which would be obtained from the release of the original microdata file. In particular, the covariance matrix and correlation matrix derived from the unmasked (original) data and microaggregated data typically differ. In this paper, we examine how the correlation between two variables can be affected through the use of microaggregation. Under microaggregation, the basic approach is to form groups of similar establishment records from the original microdata file based on some fixed definition of similarity and release group averages rather than individual components. The approach used in this paper requires the records be sorted in descending or ascending order according to the size of values of an important variable, grouped from the top of the order. The average is calculated within each group and finally the original values are replaced by their respective group averages. If there are outliers, the aggregated data may still be subject to disclosure. In that case, the outliers may need to be suppressed. In this paper, however, it will be assumed that the data does not get suppressed.

Cramer (1964) reported that the correlation obtained from the microaggregated data is always higher than that from the corresponding unmasked data. Contrary to Cramer's assertion, it is shown in this paper that the correlation obtained from the microaggregated data can be lower than that obtained from the corresponding unmasked data. A theorem and a corollary are given showing the condition under which the correlation from the microaggregated data is lower.

In addition it will be shown that:

i) the simple average of the microaggregated data based on equal or unequal subgroup size is the same as that of the unmasked;
ii) both of the above means are unbiased; and
iii) there is a form of a weighted mean which is also unbiased.

II. Properties of the Microaggregated Estimates

II.1 Notation

Let

N be the population size;

n be the total sample size;

k_i (or k) the i^{th} subgroup size;

g the number of groups

x_{ij} the variable of interest in the j^{th} observation in the i^{th} subgroup, $i=1,2,\ldots,g$; $j=1,2,\ldots,k_i$ (or k).

$\sum_{i=1}^{g} k_i = n$ or $gk = n$;

$\bar{x}_{i.}^{e} = \frac{\sum_{j=1}^{k} x_{ij}}{k}$ the sample mean of the i^{th} group based on the equal group size;

$\bar{x}_{i.}^{u} = \frac{\sum_{j=1}^{k_i} x_{ij}}{k_i}$ the sample mean of the i^{th} group based on the unequal group size;

$\bar{\bar{x}}_{..}^{e} = \frac{\sum_{i=1}^{n} \bar{x}_{i.}^{e}}{n} = \frac{\sum_{i=1}^{g} \bar{x}_{i.}^{e}}{g}$ the overall mean of the sample based on the equal group size;

$\bar{\bar{x}}_{..}^{u} = \frac{\sum_{i=1}^{n} \bar{x}_{i.}^{u}}{n} = \frac{\sum_{i=1}^{g} k_i \bar{x}_{i.}^{u}}{n}$ the overall sample mean based on the variable group size;

$\bar{\bar{x}}_{..}^{t}$ the overall sample mean of the unmasked data, i.e.,

$\bar{\bar{x}}_{..}^{t} = \frac{1}{gk} \sum_{i=1}^{g} \sum_{j=1}^{k} x_{ij}$ or $\bar{\bar{x}}_{..}^{t} = \frac{1}{n} \sum_{i=1}^{g} \sum_{j=1}^{k_i} x_{ij}$

π_{ij} the probability that the ij^{th} unit selected in sample, and

w_{ij} $= 1/\pi_{ij}$, i.e., weight assigned to the ij^{th} unit.

Note $\sum_{i} \sum_{j} w_{ij} = N$

II.2 Properties of the Sample Mean of the Microaggregated Data

Property 1. The sample mean of the microaggregated data based on the equal subgroup size ($\bar{\bar{x}}^{e}_{..}$) is identical with the sample mean of the unmasked data ($\bar{\bar{x}}^{t}_{..}$) so is the mean based on the unequal size ($\bar{\bar{x}}^{u}_{..}$).

Proof. We are to prove that

$$\bar{\bar{x}}^{e}_{..} = \bar{\bar{x}}^{t}_{..} \quad \text{and} \quad \bar{\bar{x}}^{u}_{..} = \bar{\bar{x}}^{t}_{..}$$

$$\bar{\bar{x}}^{t}_{..} = \frac{1}{gk} \sum_{i=1}^{g} \sum_{j=1}^{k} x_{ij} = \frac{1}{g} \sum_{i=1}^{g} \left(\sum_{j=1}^{k} x_{ij}/k \right) \quad (1)$$

However,

$$\bar{\bar{x}}^{e}_{..} = \frac{1}{gk} \sum_{i=1}^{g} k \, \bar{x}^{e}_{i.} = \frac{1}{g} \sum_{i=1}^{g} \left(\sum_{j=1}^{k} x_{ij}/k \right) \quad (2)$$

Hence $\bar{\bar{x}}^{e}_{..} = \bar{\bar{x}}^{t}_{..}$

Now the sample total from the microaggregation based on the unequal group size is

$$\sum_{i=1}^{g} k_i \bar{x}^{u}_{i.} = \sum_{i=1}^{g} \sum_{j=1}^{k_i} x_{ij} \quad (3)$$

Note that the right hand side of equation (3) is the same as the sample total of the unmasked data. Since the denominator for calculating the sample mean of this microaggregated data is

$$\sum_{i=1}^{g} k_i = n, \quad \text{thus,} \quad \bar{\bar{x}}^{u}_{..} \text{ is equal to } \bar{\bar{x}}^{t}_{..} .$$

Property 2. Assuming that x_{ij} is independent, identically distributed, the simple average of the microaggregated data is unbaised regardless of equal or unequal subgroup size.

Proof. If $\bar{\bar{x}}^{t}$ is unbiased, $\bar{\bar{x}}^{e}$ and $\bar{\bar{x}}^{u}$ are also unbiased, since all three are essentially the same.

We define two forms of the weighted mean.

The weighted subgroup mean for the i^{th} group is defined to be

$$\bar{x}^{w}_{i.} = \sum_{j} x_{ij} w_{ij} / \sum_{j} w_{ij}.$$

The overall weighted mean, i,e., mean of the subgroup means, can be calculated in the following two ways for the case of equal subgroup sizes:

i) $$\bar{x}^{(1)} = \sum_{i}^{g} \bar{x}^{w}_{i.}/g$$

ii) $$\bar{x}^{(2)} = \sum_{i}^{g} (\bar{x}^{w}_{i.}) \sum_{j} w_{ij}/N,$$

where $N = \sum_{i} \sum_{j} w_{ij}$.

Assuming that the subgroup used for microaggregation has nothing to do with the sample selection (original sample selection) and the size of the subgroup is fixed, we have the following property.

Property 3. $\bar{x}^{(1)}$ is biased, but $\bar{x}^{(2)}$ is not

Proof

Let $w_{i.} = \sum_{j} w_{ij}$.

Then

$$E(\bar{x}_1^{(1)}) = E\left(\sum_{i}^{n} \sum_{j} \frac{x_{ij} w_{ij}}{g \, w_{i.}}\right)$$

$$= \sum_{i}^{n} \sum_{j} \frac{x_{ij} w_j}{g \, w_{i.}} \cdot \frac{1}{w_j}$$

$$= \sum^{N} \sum x_{ij}/(g \, w_{i.})$$

Note that $\sum_{i}^{N} \sum_{j} x_{ij}$ is the population total but $g \sum_{j} w_{ij} \neq N$.

Hence $\bar{x}^{(1)}$ is biased.

$\bar{x}^{(2)}$ can be expressed as

$$\sum_{i}^{n} \sum_{j} x_{ij} w_{ij}/N$$

Hence

$$E(\bar{x}^{(2)}) = \sum_{i}^{N} \sum_{j} \frac{x_{ij} w_j}{N} \cdot \frac{1}{w_j}$$

$$= \sum_{i}^{N} \sum_{j} x_{ij}/N = \mu.$$

Thus $\bar{x}^{(2)}$ is unbiased.

In case of unequal subgroup sizes, define for $\bar{x}^{(2)}$ the weight $\bar{w}_{i.}/N$, where $\bar{w}_{i.} = \sum_{i=1}^{k_i} w_{ij}/k_i$.

Then $$\bar{x}_u^{(2)} = \sum_{i=1}^{n} \bar{x}^{w}_{i.} \, \bar{w}_{i.}/N$$

$$= \sum^{g} \bar{x}^{w}_{i.} \, w_{i.}/N$$

which is an unbiased estimator.

Note that for equal subgroups, $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ were defined as means of g values, rather than n values. However, they really are means of n values. Since n is an integer multiple (k) of g and each of g values repeats k times, $\bar{x}^{(1)}$ and $\bar{x}^{(2)}$ can be expressed as a means of g values.

It should be noted that the weighted average is more appropriate to use than the simple average, since the data is usually collected from an unequal probability sample.

We have seen that $\bar{\bar{x}}^{e}_{..} = \bar{\bar{x}}^{u}_{..} = \bar{\bar{x}}^{t}_{..}$, where $\bar{\bar{x}}^{t}_{..}$ is the overall sample mean. In the following we use $\bar{\bar{x}}_{..}$ to indicate the above means.

Property 4. Variance of the unweighted microaggregated data is no greater than that of the unweighted unmasked data.

Proof

$$(n-1)v(x) = \sum_i \sum_j (x_{ij} - \bar{\bar{x}}_{..})^2$$

$$= \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 + \sum_i \sum_j (\bar{x}_{i.} - \bar{\bar{x}}_{..})^2 \qquad (4)$$

Note that in the above $\bar{x}_{i.}$ could be either

$$\bar{x}^{e}_{i.} \text{ or } \bar{x}^{u}_{i.} .$$

From equation (4)

$$v(x) = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2}{n-1} + \frac{\sum_i \sum_j (\bar{x}_{i.} - \bar{\bar{x}}_{..})^2}{n-1} \qquad (5)$$

The second term of the right-hand side of the above equation is the sample variance of the microaggregated data. Since the first term of the right-hand side of equation (5) is non-negative, the sample variance of the microaggregated data is always less than or equal to that of the unmasked data (v(x) of equation (5)). Thus, if the sample variance of unmasked data is unbiased, the variance of the microaggregated data is biased. The equality holds if and only if every observation within a subgroup is identical, i.e., $x_{ij} = \bar{x}_{i.}$, $\forall$ i. Note that this phenomenon was observed by other researchers (see Strudler, et al page 379 and Spruill and Gastwirth page 615).

II.3 Correlation between two microaggregated variables

Cramer (1964) claimed to prove that the correlation between the two microaggregated variables is higher than the correlation between the corresponding unmasked variables. His claim is based on the assumptions that i) the within-group sum of squares is close to 0 and ii) the simple regression coefficient fitted on the microaggregated data is close to that fitted on the unmasked data (see Cramer pp. 237-241). However, both assumptions are not valid in general and thus the proof is not correct. In this section, it will be shown that the correlation obtained from the microaggregated data can be, in some cases, lower than the corresponding correlation from the unmasked data. This is true when there is a non-linear trend in the data.

Let the correlation between two unmasked variables ($Corr_t$)

$$\frac{\sum\sum(x_{ij}-\bar{\bar{x}}_{..})(y_{ij}-\bar{\bar{y}}_{..})}{\{\sum\sum(x_{ij}-\bar{\bar{x}}_{..})^2\sum\sum(y_{ij}-\bar{\bar{y}}_{..})^2\}^{1/2}}$$

$$= \{\sum\sum(\bar{x}_{i.}-\bar{\bar{x}}_{..})(\bar{y}_{i.}-\bar{\bar{y}}_{..}) + \sum\sum(x_{ij}-\bar{x}_{i.})(y_{ij}-\bar{y}_{i.})\} /$$

$$\{[\sum\sum(\bar{x}_{i.}-\bar{\bar{x}}_{i.})^2 + \sum\sum(x_{ij}-\bar{x}_{i.})^2][\sum\sum(\bar{y}_{i.}-\bar{\bar{y}}_{..})^2$$

$$+ \sum\sum(y_{ij}-\bar{y}_{i.})^2]\}^{1/2} \qquad (6)$$

Define

$$ssx_b = \sum\sum(\bar{x}_{i.} - \bar{\bar{x}}_{..})^2,$$

$$ssx_w = \sum\sum(x_{ij} - \bar{x}_{i.})^2 ,$$

$$ssxy_b = \sum\sum(\bar{x}_{i.} - \bar{\bar{x}}_{..})(\bar{y}_{i.} - \bar{\bar{y}}_{..}) ,$$

and

$$ssxy_w = \sum\sum(x_{ij} - \bar{x}_{i.})(y_{ij} - \bar{y}_{i.}) .$$

We define ssy_b and ssy_w similar to ssx_b and ssx_w. Note that in the above ssx_b is the between-group sum of squares of x and ssx_w is the within-group sum of squares of x. Similarly, $ssxy_b$ and $ssxy_w$ are the between-group and within-group cross-products respectively.

Using the above, $Corr_t$ in (6) can be reexpressed as

$$\frac{ssxy_b + ssxy_w}{(ssx_b ssy_b + ssx_w ssy_w + ssx_b ssy_w + ssx_w ssy_b)^{1/2}} \qquad (7)$$

Define

$$Corr_b = \frac{ssxy_b}{(ssx_b ssy_b)^{1/2}}$$

and

$$Corr_w = \frac{ssxy_w}{(ssx_w ssy_w)^{1/2}}$$

Note that $Corr_b$ is the correlation for the microaggregated data.

Theorem:

If $Corr_w^2 - Corr_b^2 > Corr_b^2(\frac{ssx_b}{ssx_w} + \frac{ssy_b}{ssy_w}) - 2\, Corr_b Corr_w(\frac{ssx_b}{ssx_w}\frac{ssy_b}{ssy_w})^{1/2}$, then the squared correlation from the microaggregate data ($Corr_b^2$) is lower than the squared correlation from the unmasked ($Corr_t^2$)

Proof:

By squaring expression (7), we obtain

$$Corr_t^2 = \frac{ssxy_b^2 + ssxy_w^2 + 2\, ssxy_b ssxy_w}{ssx_b ssy_b + ssx_w ssy_w + ssx_b ssy_w + ssx_w ssy_b} .$$

Note that in the above, only the first terms in the numerator and denominator constitute $Corr_b^2$. Let the extra terms in the numerator and denominator of the above be denoted by a_1 and a_2, respectively, i.e.,

$$a_1 = ssxy_w^2 + 2\, ssxy_b ssxy_w$$

and

$$a_2 = ssx_w ssy_w + ssx_b ssy_w + ssx_w ssy_b .$$

Thus, comparing $Corr_b^2$ with $Corr_t^2$ means comparing

$$\frac{ssxy_b^2}{ssx_b ssy_b} \quad \text{with} \quad \frac{ssxy_b^2 + a_1}{ssx_b ssy_b + a_2}$$

Simple algebra on the above two expressions shows that

if $a_1 > a_2$ $Corr_b^2$,

$Corr_t^2 > Corr_b^2$.

Replacing a_1 and a_2 with the original expressions, dividing the resulting inequality by $ssx_w ssy_w$, and more manipulation of some terms render the theorem.

Corollary:

If $Corr_w > Corr_b > 0$ and

$$Corr_w^2 - Corr_b^2 > Corr_b^2 \left[\left(\frac{ssx_b}{ssx_w}\right)^{1/2} - \left(\frac{ssy_b}{ssy_w}\right)^{1/2}\right]^2$$

the squared correlation from the microaggregate data is lower than that from the unmasked data.

Proof: From Theorem the right hand side of the inequality is smaller than

$$Corr_b^2 \left[\left(\frac{ssx_b}{ssx_w}\right)^{1/2} - \left(\frac{ssy_b}{ssy_w}\right)^{1/2}\right]^2$$

Thus the conditions of the theorem are satisfied.

In the following, some examples are given in which $Corr_t > Corr_b$.

Example 1:

x	1	2	3	4	5	6	7	8	9
y	10	11	12	16	18	20	13	14	15

In this example, three records each beginning from the left were microaggregated. The left hand side (LHS) of the inequality in the theorem = .70646. The right nand side (RHS) of the inequality = -3.233, i.e., LHS > RHS. Thus, this data set satisfies the condition for $Corr_b^2 < Corr_t^2$. In this case, $Corr_b$ = .4271 and $Corr_t$ = .4872. Note that $Corr_w$ = .9428 and $Corr_b$ = .4274, thus $Corr_w > Corr_b > 0$.

Note the nonlinearity in the middle of this data set. This nonlinearity caused $Corr_b^2 < Corr_t^2$.

Example 2:

This example has less nonlinearity in the middle of the data as compared with the data in example 1. Hence, the difference between $Corr_t$ and $Corr_w$ is less conspicuous than in example 1.

x	1	2	3	4	5	6	7	8	9
y	10	11	12	16	17	18	13	14	15

Again in this example, three records each beginning from the left were microaggregated. Note that only the middle two numbers (17,18) of y are different from those in example 1 (18,20). In this case

LHS = .75
RHS = -4.5
$Corr_w$ = 1.00
$Corr_b$ = .50

Hence LHS > RHS and also $Corr_w > Corr_b$. There would be, therefore, no doubt that $Corr_t^2 > Corr_b^2$. That is, from the data, $Corr_t$ = .55 and $Corr_b$ = .50.

Example 3:

This example has n = 12 and again k = 3 was used for microaggregation.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	10	11	12	16	17	18	19	20	21	13	14	15

LHS = .84
RHS = -8.1
$Corr_w$ = 1.00
$Corr_b$ = .40

Again, both conditions LHS > RHS and $Corr_w > Corr_b$ are satisfied. In actuality,

$Corr_t$ = .4336 and $Corr_b$ = .4000.

In the above, piecewise linear regression can be fitted.

Example 4:

This example slightly rearranges the y values in example 3. Large values are now located in the left side of the middle and far right. Still $Corr_t > Corr_b$.

x	1	2	3	4	5	6	7	8	9	10	11	12
y	10	11	12	16	17	18	13	14	15	19	20	21

LHS = .36
RHS = -5.4
$Corr_w$ = 1.00
$Corr_b$ = .800

From the above, we can tell $Corr_t > Corr_b$, which is true since $Corr_t$ = .8112 and $Corr_b$ = .8000.

Example 5:

This example changes example 3 by multiplying six values in the middle by 10 which again increases the nonlinearity of the data

x	1	2	3	4	5	6	7	8	9	10	11	12
y	10	11	12	160	170	180	190	200	210	13	14	15

LHS = .5965
RHS = -4.17
$Corr_w$ = .7740
$Corr_b$ = 5.02×10^{-2}

Again the conditions for $Corr_t^2 > Corr_b^2$ are met.

$Corr_t = 6.08 \times 10^{-2}$

and

$Corr_b = 5.02 \times 10^{-2}$.

Remark 1: If $Corr_w = Corr_b$ and $\frac{ssx_b}{ssx_w} = \frac{ssy_b}{ssy_w}$, then $Corr_b = Corr_t$. This can be seen from $a_1 = a_2 Corr_b^2$.

Remark 2: If $Corr_w^2 > Corr_b^2$ but $Corr_w$ and $Corr_b$ are of opposite sign, Theorem may not apply. An example for such a case is given below.

x	1	2	3	4	5	6	7	8	9
y	18	16	14	20	18	15	14	16	19

In this case

$Corr_w$ = .2287 and $Corr_b$ = .1555.

Thus

$Corr_w^2$(= .0523) is greater

than $Corr_b^2$(= .0242).

However,

$Corr_b$(= .1555) is not lower than

$Corr_t$(= .0000).

III. Concluding Remarks

Properties of the microaggregated estimates have been investigated. It has been shown that there exist unbiased weighted and weighted averages. Contrary to Cramer's proof, it is shown in this paper that the correlation obtained from the micro-aggregated data can be lower than that obtained from the corresponding unmasked data. A theorem and a corollary are given showing the condition under which the correlation from the microaggregated data is lower. A condition for the two correlation to be equal is also given. In the case outliers are removed from the microaggregate data file the results in this paper will not hold exactly. The properties of other forms of microaggregation (see Wolf (1988)) are under investigation.

References

Cochran, W.G., Sampling Techniques, Third Edition, John Wiley and Sons, Inc. (1977)

Cramer, J.S., "Efficient Grouping, Regression and Correlation in Engel Curve Analysis," Journal of the American Statistical Association, Vol. 59, 233-250 (1964).

Govoni, J.P. and Waite, P.J., "Development of A Public Use File for Manufacturing," American Statistical Association 1985 Proceedings of the Business and Economic Statistica Section, 300-308 (1985).

Praise, S.J. and Aichison, J., "The Grouping of Observations in Regression Analysis," Review of the International Statistical Insititute, Vol. 22, 1-22 (1954).

Spruill, N.L., "Protecting Confidentiality of Business Microdata by Masking," The Public Research Institute, 1984.

Spruill, N.L. and Gastwirth, J.L., "On the Estimation of the Correlation Coefficient From Grouped Data," Journal of the American Statistical Association, Vol. 77, 614-620 (1982)

Strudler, M., Oh, H.L. and Scheuren, F., "Protection of Taxpayer Confidentiality with Respect to the Tax Model," American Statistical Association 1986 Proceedings of the Section on Survey Research Methods, 375-381 (1986).

Wolf, M.K., "Microaggregation and Disclosure Avoidance for Economic Establishment Data," presented at the American Statistical Association meeting in New Orleans, 1988.

1/ 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.