Creating Synthetic Microdata for Higher Educational Use in Japan: Reproduction of Distribution Type based on the Descriptive Statistics

Kiyomi Shirakawa[*], Yutaka Abe[**] and Shinsuke Ito[***]

[*] Hitotsubashi University / National Statistics Center, 2-1 Naka, Kunitachi-shi, Tokyo 186-8603, Japan, kshirakawa@ier.hit-u.ac.jp
[**] Hitotsubashi University, 2-1 Naka, Kunitachi-shi, Tokyo 186-8603, Japan, y-abe@ier.hit-u.ac.jp
[***] Chuo University, 742-1 Higashinakano, Hachioji-shi, Tokyo 192-0393, Japan, ssitoh@tamacc.chuo-u.ac.jp

Abstract: When creating synthetic microdata in Japan, the values from result tables are used in order to remove links to individual data. The result tables of conventional official statistics do not allow the generation of random numbers for reproducing the individual data. Therefore, the National Statistics Center has created pseudo-individual data on a trial basis using the 2004 National Survey of Family Income and Expenditure.
Although mean, variance, and correlation coefficient in the original data were reproduced in the synthetic microdata created, the trial did not include the creation of completely synthetic microdata from the result tables, and the reproduction of the distribution was not taken into account.
In this study, a method for generating random numbers with a distribution close to that of the original data was tested. It is called 'Academic Use File'. The random numbers were generated completely from the values contained in the result tables. In addition, this test took into account the Anscombe's quartet, and the sensitivity rule. As a result, based on the numerical values of the result tables, it was possible to introduce the closest approach to the distribution type of the original data.

1 Introduction

When creating synthetic microdata in Japan, the values from result tables are used in order to remove links to individual data in order to comply with Japanese legal requirements. Therefore, the National Statistics Center has created pseudo-individual data on a trial basis using the 2004 National Survey of Family Income and Expenditure, where the mean, variance, and correlation coefficient in the original data were reproduced in the synthetic microdata created. Here, synthetic microdata is used to refer to microdata that can be accessed without an application and used without restrictions. However, this trial did not include sufficient information about statistical tables such as kurtosis and skewness and therefore did not allow the creation of completely synthetic microdata based on statistical tables. The distribution type of the synthetic microdata was therefore not taken into account in reproducing the distribution type of the original data.

In this research, we tested a method for generating random numbers with a distribution close to that of the original data. Random numbers were generated completely from the values posted in the result tables. For this work, Anscombe's quartet was taken into

account. Also, based on the numerical values in the result tables, we aimed to establish the closest approach to the distribution type of the original data.

2 Problems with and Improvements to Synthetic Microdata

2.1 Applicability of Microaggregation to Synthetic Microdata

Microaggregation is one of the disclosure limitation methods adopted for official microdata. Microaggregation entails dividing the individual records into groups larger than a threshold k and replacing the records with common values as measures of the central tendency (e.g., the mean) within each group. The method of microaggregation was developed based on research by Defays and Anwar (1998), Domingo-Ferrer and Mateo-Sanz (2002) and others.

Ito et al. (2008) and Ito (2009) applied the methodology of microaggregation to Japanese official microdata, identified the applicability of microaggregation to synthetic microdata, and evaluated the effectiveness of microaggregation for individual data from the National Survey of Family Income and Expenditure. These studies were the first in Japan to advocate methods using multi-dimensional cross-tabulation to create microaggregated data that closely resembles individual data. The proposed method of microaggregation is as follows. In the first step, records with common values for all types of qualitative attributes based on multi-dimensional cross-tabulation were created. In the second step, records with common values for qualitative attributes were sorted and divided into groups larger than a specific threshold, and the value of each quantitative attribute for records was replaced with an average value within each group.

Microaggregation is generally applied to the quantitative attributes contained in microdata. For such attributes, if the records containing a common attribute value are grouped for every target qualitative attribute and these attribute values are viewed as being replaced with representative values for the group, then grouping of records related to qualitative attribute values can also be positioned as a form of microaggregation. In this case, the microaggregated data are considered to be the set of the same qualitative attribute values within a particular group and the corresponding set of records that contain the mean values of the quantitative attributes. Although this kind of microaggregated data can be viewed as data that conform to individual data consisting of a set of qualitative attribute values and a set of mean values of quantitative attributes, the set of attribute values of each of the records can be positioned as only aggregate values.

Although cross-tabulation tables can be created by the grouped target qualitative attributes, the frequency of the designated cells within a cross-tabulation table matches the number of records within the corresponding group in the microaggregated data. This means that the number of qualitative attributes used for the grouping of records increases as the dimensionality of the cross-tabulation table increases. By expanding on this methodology, we can define "hyper-multidimensional cross-tabulation tables"

which are "n-dimensional cross-tabulation tables created by tabulating the set of all attributes of the individual data" (Ito (2008)), and we can logically construct a set of microaggregated data that characterize the set of records having a correspondence with the cells contained in the cross-tabulation table. Note that hyper-multidimensional cross-tabulation tables include all dimensions of cross-tabulation tables from 1 to n dimensions (Bethlehem et al. 1990, Höhne 2003). This means that various dimensions of cross-tabulation tables can be created for setting the hyper-multidimensional tabulation tables and can serve as the basis for creating synthetic microdata in the framework of hyper-multidimensional cross-tabulation tables.
As mentioned above, the characteristics of microaggregation are that the records contained in the individual data are grouped into a set of records with a threshold value k, and the individual attribute values in the records of the group are replaced with a representative value such as the mean value. This indicates that the number of records that exist within the set of records with common values for qualitative attributes has a correspondence relationship with the frequency of cells in the hyper-multidimensional cross-tabulation tables created with the same set of attributes. Therefore, once the lower limit on the number of records contained in the set of records with common values for qualitative attributes has been set, this determines the threshold value for the frequency of cells contained in the hyper-multidimensional cross-tabulation table. When the threshold k is set, a cross-tabulation table can be created by appropriately selecting the combination of attributes from the set of attributes that form the aggregation items in the hyper-multidimensional cross-tabulation table in such a way that no cells contained in the hyper-multidimensional cross-tabulation table are zero and all the cells have a frequency of at least k. Furthermore, if cells with a frequency less than the threshold k exist in the hyper-multidimensional cross-tabulation table, then it is possible to perform grouping into the set of records with common values for qualitative attributes with the threshold k or higher by performing processing based on “unknowns” in the group of attributes of records corresponding to those cells.
By doing this the creation of data that conforms to individual data based on hyper-multidimensional cross-tabulation data can be methodologically positioned within the microaggregation framework. This demonstrates that microaggregation forms a logical foundation in the method of creating synthetic microdata for education.

2.2 Creating Synthetic Microdata

Synthetic microdata for public Japanese microdata were created based on the methodology of microaggregation. This section describes how the synthetic microdata were created using multi-dimensional tabulation, in reference to Section 3 of Makita et al. (2013). The detailed process for creating synthetic microdata is as follows.

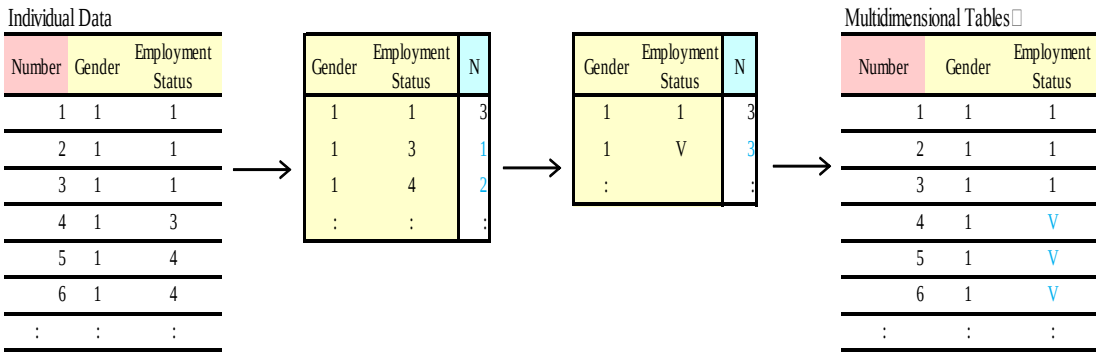
First, quantitative and qualitative attributes to be contained in the synthetic microdata were selected. Second, records with common values for qualitative attributes were sorted into groups with a minimum size of 3. Third, tables were created in order to

generate multivariate lognormal random numbers and records for which the values for some quantitative attributes were 0. This process allowed the creation of synthetic microdata with characteristics similar to those of the original microdata (Makita et al. (2013, p. 2)).
Figure 1 and Figure 2 present the detailed process of creating the synthetic microdata, as described below.

(1) Qualitative attributes were selected from the multi-dimensional statistical tables compiled based on the original microdata. Specifically, 14 qualitative attributes were selected based on the survey items used most frequently by researchers, including gender, age, and employment status. In addition, 184 quantitative attributes were selected, including yearly household income and monthly household expenditures.

(2) Records with common values for qualitative attributes were sorted into groups with a minimum size of 3. For records that have common values for some qualitative attributes and that refer to groups with a size of 1 or 2, values for the other qualitative attributes were transformed to 'unknown' (V) in order to create groups with a minimum size of 3.

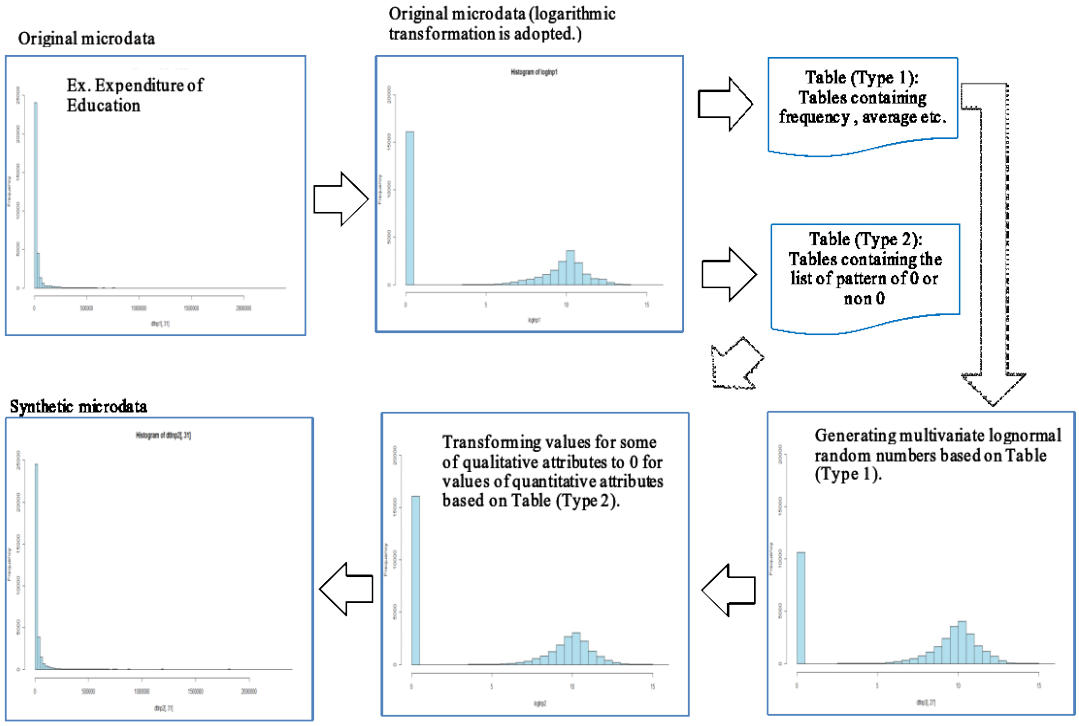
(3) Two types of tables were created in order to generate 1) multivariate lognormal random numbers and 2) records with negative values for some quantitative attributes. Tables of 'Type 1' contain frequency, mean, variance, and covariance of quantitative attributes not including 0. The records on which these tables are based were classified by qualitative attribute in order to generate multivariate lognormal random numbers. Tables of 'Type 2' are tables created by sorting records based on whether values for quantitative attributes are 0 or not 0, and on this basis, the values for some quantitative attributes in the records were transformed to 0 (Makita et al. (2013, p.3)).



Note: "V" stands for "unknown".
Source: Makita et al. (2013).

Figure 1. Processing records with common values for qualitative attributes into groups with a minimum size of 3.

To create the synthetic microdata, logarithmic transformation was applied to the original microdata items. Then, multivariate lognormal random numbers were created based on the above two types of tables, and the values for some quantitative attributes were transformed to 0. As a final step, exponential transformation was conducted. Figure 1 illustrates how to process records with common values for qualitative attributes into groups with a minimum size of 3. Figure 2 shows the creation of the synthetic microdata and compares the frequency of the synthetic microdata with that of the original microdata.



Source: Makita et al. (2013).

Figure 2. Creation of the synthetic microdata and comparison between the frequency of the synthetic microdata and that of the original microdata.

2.3 Problems in Creating Synthetic Microdata

This section discusses problems with the synthetic microdata.

(1) All variables were subjected to exponential transformation in units of cells in the result table.
Table 1, which was created from the synthetic microdata, contains several standard deviations that are too large.

Table 1. Indicators of living expenditures and food in workers' households (household size: 4 persons).

Number of earners	Structure of dwelling	Frequency	Living expenditure			Food		
			Mean	SD	C.V.	Mean	SD	C.V.
One person		4,132	302,492.8	148,598.9	0.491	71,009.0	25,089.5	0.353
	Wooden	1,436	300,390.3	170,211.4	0.567	71,018.5	24,187.6	0.341
	Wooden with fore roof	501	298,961.0	125,682.9	0.420	73,507.3	24,947.7	0.339
	Ferro-concrete	1,624	306,947.4	131,895.0	0.430	69,873.1	25,844.2	0.370
	Unknown	571	298,209.7	153,651.1	0.515	72,024.1	25,125.1	0.349
Two persons		4,201	346,195.7	215,911.7	0.624	78,209.1	25,288.1	0.323
	Wooden	1,962	346,980.3	172,673.2	0.498	78,961.7	24,233.5	0.307
	Wooden with fore roof	558	356,021.5	160,579.8	0.451	81,039.4	24,628.2	0.304
	Ferro-concrete	1,120	353,093.9	313,837.8	0.889	76,860.8	26,250.7	0.342
	Others	3	260,759.8	37,924.3	0.145	72,733.1	5,358.9	0.074
	Unknown	558	320,224.5	148,230.3	0.463	75,468.5	27,241.1	0.361

(2) Correlation coefficients (numerical) between all variables were reproduced.
From Table 2, several correlation coefficients were too small. This was because correlation coefficients between uncorrelated variables were also reproduced.

Table 2. Correlation coefficients of each variable.

	Living expenditure	Food	Housing
Living expenditure	1.00	0.5	0.28
Food	0.43	1.00	-0.03
Housing	0.28	-0.06	1.00

Top half: original data; bottom half: synthetic microdata.

(3) Qualitative attributes of groups having a frequency (size) of 1 or 2 were transformed to "Unknown" (V) or deleted.
The information loss when using this method was too large. Furthermore, the variations within the groups were too large to merge qualitative attributes between different groups.

2.4 Correcting the Trial Synthetic Microdata

This section presents corrections for approximating the distribution types of the original data.

(1) Select the transformation method (logarithmic transformation, exponential transformation, square-root transformation, reciprocal transformation) based on the original distribution type (normal, bimodal, uniform, etc.). Note that exponential transformation was used for all transformations when creating the synthetic microdata here.

$$f(x|\lambda) = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log x & (\lambda = 0) \end{cases}$$

$\lambda = 0$ logarithmic transformation
$\lambda = 0.5$ square-root transformation
$\lambda = -1$ reciprocal transformation
$\lambda = 1$ linear transformation

(2) Detect non-correlations for each variable.
Correlation coefficients are reproduced between only variables that have a correlation relationship:

$$\mathrm{T}(r,0) = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$$ r: correlation coefficient

The detection results are confirmed by using the two-tailed Student's t distribution.

(3) Qualitative attributes in groups with a size of 1 or 2 are merged into a group that has a minimum size of 3 in the upper hierarchical level.
Note that Anscombe's quartet shows four groups that have the same frequency, mean, standard deviation, and regression model parameters. However, the distribution types of these groups are different.

Table 3. Examples of numerical values for Anscombe's quartet.

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Sample variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	y = 3.00 + 0.500x (to 2 and 3 decimal places, respectively)

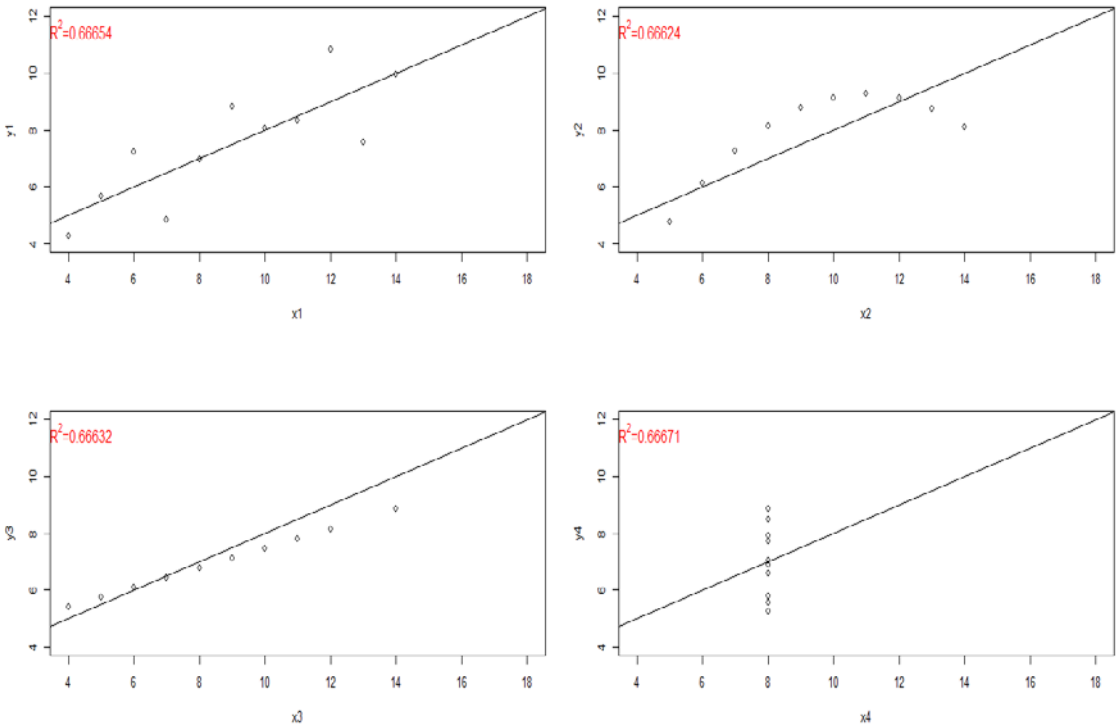


Figure 3. Scatter plots of numerical examples for Anscombe's quartet.

This indicates that second moments can be reproduced based on the mean and standard deviation. However, it also indicates that third and fourth moments (skewness and kurtosis) cannot be reproduced. More specifically, we can see that the numerical values of the kurtosis and skewness differ from those of the original microdata.
To resolve these problems, it is necessary for the numerical values of the third and fourth moments to approximate those of the original microdata. The specific indicators are frequency, mean, standard deviation, kurtosis, and skewness. Furthermore, to create the synthetic microdata (here, pseudo-microdata created by microaggregation) based on multivariate normal random numbers, λ in the Box-Cox transformation is required in order to change the distribution type of the original data into a standard distribution. Note that these indicators are the minimum indicators for reproducing the original microdata, and are not absolute indicators.

3 Creating Academic Use File

(1) Create microdata based on kurtosis and skewness

After creating several multivariate normal random numbers, a random number that approximates the kurtosis and skewness of the original microdata was selected.

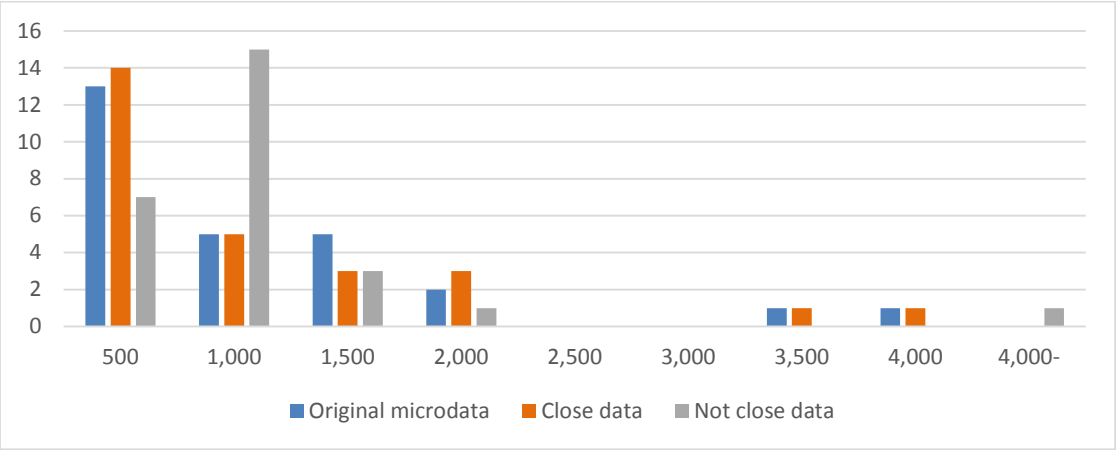


Figure 4. Differences of kurtosis and skewness.

From Figure 4, the synthetic microdata have approximately the same kurtosis and skewness as the original microdata. This figure shows that the contributions to kurtosis and skewness are clear. We call this new synthetic microdata 'Academic Use File', and call existing synthetic microdata 'Public Use File'.

From Table 4, the value of λ is 0. In this case, the logarithmic transformation is optimal in the Box-Cox transformation.

Table 4. Original microdata and transformed indicators for each transformation.

	Original data	Log2 transformation	Natural lognormal transformation	Square-root transformation	Reciprocal transformation
Mean	861.370	9.139	6.335	26.451	2.651
SD	882.057	1.363	0.945	12.960	2.548
Kurtosis	4.004	*-0.448*	*-0.448*	0.974	4.185
Skewness	2.002	*0.107*	*0.107*	1.115	1.943
Frequency	27				
λ	-0.047 (λ = 0)				

(2) Create microdata based on the two tabulation tables of the basic table and details table

In this research, we created additional academic use file by treating the first set of synthetic microdata as original microdata. The sample data (Table 5), the basic table (Table 6), and the details table (Table 7) are shown in order to explain the method for

creating the academic use file. Note that apart from the corrections, the method for creating the academic use file is the same as for creating the first set of synthetic microdata.

Table 5. Sample data (individual data).

Group No.	A	B	C	D	E	F	Living expenditure	Food	Housing
1	2	1	1	2	5	1	125,503.5	29,496.1	2,171.6
	2	1	1	2	5	1	255,675.9	25,806.2	
	2	1	1	2	5	1	175,320.4	38,278.2	
2	2	1	1	3	6	1	181,085.6	74,122.1	
	2	1	1	3	6	1	124,471.0	33,256.8	329.6
	2	1	1	3	6	1	145,717.7	46,992.8	
3	2	1	1	3	7	1	319,114.3	113,177.1	263.3
	2	1	1	3	7	1	253,685.2	67,253.6	341.4
	2	1	1	3	7	1	236,447.6	61,129.8	
4	3	1	1	1	5	1	137,315.3	27,050.1	9,256.0
	3	1	1	1	5	1	253,393.7	47,205.6	
	3	1	1	1	5	1	232,141.8	52,259.6	
	3	1	1	1	5	1	214,540.4	54,920.9	
5	3	1	1	1	6	1	234,151.4	74,993.0	
	3	1	1	1	6	1	278,431.0	78,916.1	110.7
	3	1	1	1	6	1	197,180.8	72,909.6	
6	3	1	1	2	5	1	118,895.1	48,821.6	408.5
	3	1	1	2	5	1	130,482.8	47,798.5	
	3	1	1	2	5	1	147,969.1	50,277.9	309.0
	3	1	1	2	5	1	150,973.7	48,291.0	

A: 5-year age groups; B: employment/unemployed; C: company classification; D: company size; E: industry code; F: occupation code

Table 6. Basic table (matches with original mean and standard deviation, approximate correlation coefficients for each variable).

	Living expenditure	Food	Housing
Mean	195,624.8	54,647.8	1,648.8
SD	59,892.6	21,218.1	3,144.4
Kurtosis	-1.004164	1.628974	6.918601
Skewness	0.346305	0.992579	2.605260
Frequency	20	20	8

Correlation coefficients	Living expenditure	Food	Housing
Living expenditure	1		
Food	0.643	1	
Housing	-0.335	-0.489	1

Table 7. Details table (means and standard deviations for creating synthetic microdata for multidimensional cross fields).

Groups	Living expenditure			Food		
	Frequency	Mean	SD	Frequency	Mean	SD
1	3	185,499.9	65,680.5	3	31,193.5	6,406.9
2	3	150,424.8	28,599.3	3	51,457.2	20,795.2
3	3	269,749.0	43,611.7	3	80,520.1	28,447.0
4	4	209,347.8	50,580.8	4	45,359.0	12,618.4
5	3	236,587.8	40,679.9	3	75,606.2	3,049.8
6	4	137,080.2	15,119.7	4	48,797.2	1,071.9

In this research, we created academic use file based on the correction described in Section 2.4 above.

We employed the following two tables:

Basic table: Frequency, mean, standard deviation, kurtosis, skewness, and λ in Box-Cox transformation.
Details table: Frequency, mean, and standard deviation

Several multivariate normal random numbers were generated based on the mean and standard deviation from the basic table. Next, we selected random numbers that are near the kurtosis and skewness of the original microdata. From this, we performed transformation based on non-correlation detection and the λ in the Box-Cox transformation. Finally, we replaced the random numbers we have been working with up to now with the mean and standard deviation within each group in the details table. By doing this, the numerical values of each of the variables in the synthetic microdata matched the numerical values of the variables in the details table, and we obtained multivariate microdata. Furthermore, the mean and standard deviation were the same

and the kurtosis and skewness were approximately the same at the same level of dimensionality as the basic table (number of multivariate cross fields).
Note that if there were groups of size 1 or 2 in the details table, those qualitative attributes were not transformed to unknown (V). Furthermore, those records were also not deleted. This was because groups of size 1 or 2 were merged into groups at the same level as the basic table (upper hierarchy level).

(3) Create microdata based on multivariate normal random numbers and exponential transformation
This is a method for creating trial synthetic microdata. Refer to Section 2.2 above for details. Furthermore, we also tested other methods and specifically looked at microaggregation with a threshold of 3. This is a method of sorting the values of variables in ascending order, dividing them into groups of minimum size 3, and creating synthetic microdata based on the means and standard deviations in these groups. This method is very simple and useful, but it was not suitable for creating synthetic microdata based on public statistics result tables because the multivariate variables cannot be sorted in ascending order for each variable. As a result, this method was excluded from this research.

4 Sensitivity Rules for Academic Use File

We suggested the creation of academic use file. However, if the academic use file has disclosure risk, we do not open it same as for Public Use File.
Table 8 briefly presents the most common sensitivity rules measure to assess disclosure risk. First is minimum frequency rule. It defines that a cell is considered unsafe if the cell frequency is less than pre-specified minimum frequency. Next is (n, k) rule. Cells are considered unsafe if the sum of the n largest contributions exceeds k% of the cell total. Last is p% rule. A cell is considered unsafe if the cell total minus 2 largest contributions is less than p% of the largest contribution.

Table 8. the most common sensitivity rules.

Rule	Definition : A cell is considered unsafe
Minimum frequency rule	the cell frequency is less than a pre-specified minimum frequency n (the common choice is $n = 3$).
(n, k) rule	the sum of the n largest contributions exceeds k% of the cell total X, e.g $$x_1 + x_2 + \cdots + x_n > \frac{k}{100} \cdot X$$
p% rule	the cell total X minus 2 largest contributions x_1 and x_2 is less than p% of the largest contribution, e.g. $$X - x_1 - x_2 < \frac{p}{100} \cdot x_1$$

Even so, if the academic use file has no problem with these rules, it does not mean that the microdata is safe. We consider it with trial data combinations. Each combination has N integer variables, total of variables is 100 and each variables are arranged in descending order. For example, we use $N = 20$ integer variables $x_1, x_2, \ldots, x_{20}$ combinations. The above conditions requirement $x_1 + x_2 + \cdots + x_{19} + x_{20} = 100$ and $x_1 \geq x_2 \geq \cdots \geq x_{19} \geq x_{20}$. Number of these combinations is 97,132,873 patterns.
Then we use $p\%$ rule to check safe or unsafe combinations. In Table 9 we show frequencies of unsafe combinations for each largest value in combination. Only 8,849 combinations is unsafe by $p\%$ rule. They occupy only 0.01% of the total combinations.

Table 9. Frequency of unsafe combinations with p% rule.

x_1	frequency	x_1	frequency	x_1	frequency	x_1	frequency	x_1	frequency
100	**1**	**89**	**19**	**78**	**195**	**67**	**373**	**56**	**195**
99	**1**	**88**	**30**	**77**	**195**	**66**	**373**	**55**	**139**
98	**2**	**87**	**30**	**76**	**272**	**65**	**272**	**54**	**139**
97	**2**	**86**	**45**	**75**	**272**	**64**	**272**	**53**	**139**
96	**4**	**85**	**45**	**74**	**373**	**63**	**272**	**52**	**139**
95	**4**	**84**	**67**	**73**	**373**	**62**	**272**	**51**	**139**
94	**7**	**83**	**67**	**72**	**508**	**61**	**272**	**50**	**97**
93	**7**	**82**	**97**	**71**	**508**	**60**	**195**	**49**	**95**
92	**12**	**81**	**97**	**70**	**373**	**59**	**195**	**48**	**90**
91	**12**	**80**	**139**	**69**	**373**	**58**	**195**	**46**	**52**
90	**19**	**79**	**139**	**68**	**373**	**57**	**195**	**47**	**78**
Total					**8,849**				

However, you can narrow the combinations with basic statistic. If you know a combination's standard deviation, skewness and kurtosis, there are only 16 combinations at most which has same statistics. Furthermore, if you know median or second-largest value, you can narrow down more the combinations. There are only 9 combinations at most (Table 10).

Table 10. Maximum number of combinations grouping combinations by each statistic.

StDev.	Skew.	Kurt.	Median	Intruder	N=30	N=20	difference
*					**331,258**	**223,627**	**107,631**
*	*				**873**	**550**	**323**
*	*	*			**23**	**16**	**7**
*	*	*	*		**23**	**12**	**11**
*	*	*	*	*	**22**	**9**	**13**

Each combination has a set of 3 statistics; standard deviation, skewness and kurtosis. We make groups of combinations which has same set of statistics. Then, we sort the groups by size, and count number of sets representing each groups.
For example, there exist a set of standard deviation, skewness and kurtosis, such that 16 combinations has that set. Then we found another set such that 16 combinations has that set. Thus, 16 in count column has 2 frequency.
Even if count is 2, there are only 8,524,260 frequency (Table 11). As a matter of fact, 87.7% combinations of the total has count 1. This means most of combinations are easily identified, if it is known the standard deviation, skewness and kurtosis.
Since academic use file reproduce original distribution type, the microdata has a certain amount of disclosure risk.
Thus we cannot open the data to the public and it must be appropriately limited such as registration system.

Table 11. Range of each statistic grouping by standard deviation, skewness and kurtosis (freq. N=20).

Count	freq.	Max. StDev.	Mini. StDev.	Max. Skew.	Mini. Skew.	Max. Kurt	Mini. Kurt
16	2	4.180153611	3.741657387	0.57644737	0.361706944	-0.587458267	-0.840557276
15	13	4.565315462	3.524351377	0.964551851	0.192366206	0.309376382	-1.054150134
14	64	4.565315462	3.324549831	0.955694047	0.145282975	0.530688627	-1.231641448
13	155	4.963021151	3.077935056	0.901252349	0.124964037	0.334557548	-1.371295887
12	445	5.211323703	2.901905	1.105031963	0.083177673	0.882366328	-1.325211176
11	1,153	5.380275868	2.91998558	1.35646267	0.060994516	1.329705653	-1.43192732
10	3,233	5.830951895	2.695024656	1.50255637	-0.009153874	2.780946447	-1.463372549
9	8,186	5.938279035	2.675424216	2.643593746	-0.054613964	8.856069776	-1.506246692
8	20,059	6.316228055	2.533979604	2.790656548	-0.129051837	9.651964674	-1.619289547
7	46,302	6.844129255	2.406132516	3.167347883	-0.377822461	11.83376246	-1.677127049
6	109,605	7.813618341	2.339590607	3.529878009	-0.44212357	14.2766552	-1.701512451
5	269,146	8.926601286	2.152110347	3.859596389	-0.595837348	16.21353047	-1.755565919
4	718,999	10.35679284	1.91942974	4.019587737	-0.71911404	17.17271192	-1.858246651
3	2,210,969	13.58404637	1.716790151	4.237554114	-1.141558149	18.50919755	-1.934757558
2	8,524,260	15.97366253	1.376494403	4.412088065	-1.578947368	19.62372574	-2.036823063

5 Comparison between Various Sets of Synthetic Microdata

In order to compare various sets of synthetic microdata, we selected synthetic microdata that most closely approximated the original microdata. Furthermore, we selected indicators for creating the optimal synthetic microdata. We compared the characteristics with the original data in order to establish how easy the synthetic microdata are to use. Table 12 shows various indicators for the original microdata and three sets of synthetic microdata.

The number of observation values was 20 in all of the microdata, and the means and standard deviations were also the same. Furthermore, the correlation coefficients were either the same (column numbers 3 and 4) or approximately the same (column number 2) as those of the original microdata.

Note that the correlation coefficients for all of the synthetic microdata were the same as those for the original microdata. However, because the synthetic microdata for column number 2 was transformed from the means and standard deviations in the six groups in the details table and not from the means and standard deviations in the basic table after creating the random numbers, they do not match due to variations in the values between when the random numbers were created and after transformation. In addition, the indicators for the skewness, kurtosis, maximum value, and minimum value differ between the different microdata.

The most useful microdata from the indicators in Table 12 are in column number 2. Next are those in column number 3, and finally column number 4. Note that for reference, column number 4 is the same as the trial synthetic microdata method.

Table 12. Comparison of original microdata and each set of synthetic microdata.

No.	1 Original microdata		2 Hierarchization, and kurtosis, skewness and λ of Box-Cox transformation		3 Kurtosis and skewness		4 Multivariate lognormal random numbers	
	Living expenditure	Food	Living expenditure	Food	Living expenditure	Food	Living expenditure	Food
1	125,503.5	29,496.1	110,487.8	25,143.0	107,684.0	23,459.9	133,549.9	38,559.9
2	255,675.9	25,806.2	232,691.8	37,905.5	281,880.8	56,520.4	123,716.6	42,930.1
3	175,320.4	38,278.2	213,320.2	30,531.9	254,267.3	37,419.4	152,784.8	67,263.8
4	181,085.6	74,122.1	183,430.4	75,469.1	294,589.9	112,843.9	195,764.8	8,286.1
5	124,471.0	33,256.8	134,867.6	39,568.9	193,191.6	54,363.3	202,865.8	75,558.0
6	145,717.7	46,992.8	132,976.4	39,333.7	189,242.7	53,980.3	193,003.4	70,994.2
7	319,114.3	113,177.1	242,622.5	68,472.2	151,183.6	55,303.2	191,620.1	52,311.7
8	253,685.2	67,253.6	320,055.9	113,008.5	271,338.1	79,991.4	72,773.7	13,621.6
9	236,447.6	61,129.8	246,568.6	60,079.7	157,306.9	50,650.9	201,114.6	74,899.0
10	137,315.3	27,050.1	144,192.6	32,572.9	167,431.0	36,116.3	217,530.7	60,736.0
11	253,393.7	47,205.6	267,708.8	60,344.8	270,301.8	78,246.4	297,608.7	77,464.3
12	232,141.8	52,259.6	212,050.7	37,656.3	223,946.8	43,827.9	175,993.6	71,416.6
13	214,540.4	54,920.9	213,439.1	50,862.2	225,103.2	63,861.2	297,653.0	86,400.5
14	234,151.4	74,993.0	205,595.0	73,919.1	165,972.3	49,350.6	123,197.1	31,645.5
15	278,431.0	78,916.1	282,652.7	79,126.9	249,749.1	73,474.1	277,501.6	69,910.5
16	197,180.8	72,909.6	221,515.6	73,772.7	183,281.1	48,672.3	235,221.1	58,700.6
17	118,895.1	48,821.6	127,964.3	50,240.7	115,639.3	71,059.5	182,363.2	49,433.2
18	130,482.8	47,798.5	159,328.0	48,533.5	170,231.1	38,723.5	158,939.4	45,131.8
19	147,969.1	50,277.9	133,795.5	47,660.6	125,789.2	22,188.5	212,194.2	37,995.6
20	150,973.7	48,291.0	127,232.9	48,754.2	114,366.4	42,903.1	267,100.1	59,697.3
Mean	195,624.8	54,647.8	195,624.8	54,647.8	195,624.8	54,647.8	195,624.8	54,647.8
SD	59,892.6	21,218.1	59,892.6	21,218.1	59,892.6	21,218.1	59,892.6	21,218.1
Kurtosis	-1.004164	1.628974	-0.810215	1.473853	-1.220185	1.721354	-0.212358	-0.052164
Skewness	0.346305	0.992579	0.310913	1.050568	0.160612	0.949106	0.035785	-0.709361
Correlation coefficients	0.642511		***0.689447***		0.642511		0.642511	
Maximum	319,114.3	113,177.1	320,055.9	113,008.5	294,589.9	112,843.9	297,653.0	86,400.5
Minimum	118,895.1	25,806.2	110,487.8	25,143.0	107,684.0	22,188.5	72,773.7	8,286.1

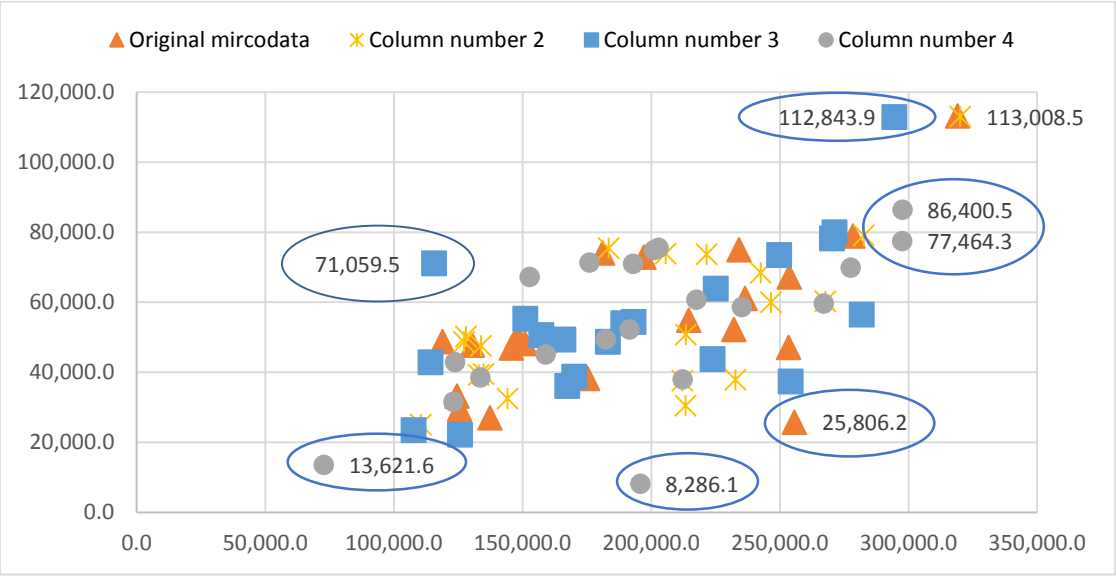


Figure 5. Scatter plots of living expenditure and food for each microdata.

From Figure 5, column number 2 approximates the original microdata, and column numbers 3 and 4 contain several outliers. This result shows that kurtosis, skewness, and Box-Cox transformation λ are useful indicators for synthetic microdata, and furthermore that transformation using the mean and standard deviation from the details table (lower hierarchical level) is required after creating the random numbers. Note that Table 13 shows an example of the result table for creating the optimal synthetic microdata.

Table 13. Example of the result table for creating academic use file.Example of the result table for creating academic use file.

Items							**Living expenditure**			**Food**		
No.	**A**	**B**	**C**	**D**	**E**	**F**	**Frequency**	**Mean**	**SD**	**Frequency**	**Mean**	**SD**
1	**2**	**1**	**1**	**2**	**5**	**1**	**3**	**185,499.9**	**65,680.5**	**3**	**31,193.5**	**6,406.9**
2	**2**	**1**	**1**	**3**			***6***	***210,086.9***	***73,208***	***6***	***65,988.7***	***27,387.3***
	2	**1**	**1**	**3**	**6**	**1**	**3**	**150,424.8**	**28,599.3**	**3**	**51,457.2**	**20,795.2**
	2	**1**	**1**	**3**	**7**	**1**	**3**	**269,749.0**	**43,611.7**	**3**	**80,520.1**	**28,447.0**
3	**3**	**1**	**1**	**1**			***7***	***221,022.1***	***45,197.7***	***7***	***58,322.1***	***18,550.2***
	3	**1**	**1**	**1**	**5**	**1**	**4**	**209,347.8**	**50,580.8**	**4**	**45,359.0**	**12,618.4**
	3	**1**	**1**	**1**	**6**	**1**	**3**	**236,587.8**	**40,679.9**	**3**	**75,606.2**	**3,049.8**
4	**3**	**1**	**1**	**2**	**5**	**1**	**4**	**137,080.2**	**15,119.7**	**4**	**48,797.2**	**1,071.9**
Mean								**195,624.8**			**54647.8**	
Standard deviation							**59,892.6**			**21218.1**		
Kurtosis							**-1.004**			**1.629**		
Skewness							**0.346**			**0.993**		
Correlation coefficients							**0.643**					
λ							**0**					

6 Conclusions and Future Outlook

In this paper, we focused on improvements to trial synthetic microdata created by the National Statistics Center for statistics education and training. The synthetic microdata created by National Statistics Center are not a duplicate of the original microdata, but rather a substitute suitable for statistics education and training. More specifically, these synthetic microdata were created by using microaggregation, which is a disclosure limitation method for public statistical microdata.

In addition, we attempted to create academic use file using several methods that adhere to this disclosure limitation method. The results show that kurtosis, skewness, and Box-Cox transformation λ are useful in addition to the frequency, mean, standard deviation, and correlation coefficient which have previously been used as indicators. However, there are no examples containing the indicators we examined in this work (kurtosis, skewness, and Box-Cox transformation λ) in Japanese public statistics result tables. In particular, λ is used only for converting the original microdata distribution type into normal distributions, and publication of the numerical value is not meaningful. Furthermore, even without knowing the value of λ, the methods for transforming the normal distributions can be limited to three. The conclusion of this paper is to take the tabulation table with the kurtosis and skewness added to the conventional indicators as a basic table. Furthermore, for correlation relationships, the correlation coefficients (numerical values) between variables are reproduced based on detection of non-correlations. Transformations to the frequency, mean, and standard deviation in each group are based on a details table (multi-dimensional cross fields). By doing this, it is possible to create academic use file that approximate the original microdata.

Problems for the future are deciding number of cross fields (dimensionality) of the basic table and details table and the style (indicators to tabulate) of the result table according to the statistical fields in the public survey. The reason is that new trials will be necessary if there is a lack of indicators based on this conclusion. Furthermore, we aim to expand this work to the creation and correction of synthetic microdata for other surveys. In the future, we will create synthetic microdata for several surveys and establish a method for creating synthetic microdata in Japan.

References

Anscombe, F.J. (1973), “Graphs in Statistical Analysis” *American Statistician*, 17-21.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990) “Disclosure Control of Microdata”, *Journal of the American Statistical Association*, Vol. 85, No. 409 pp.38-45.

Defays, D. and Anwar, M.N. (1998) “Masking Microdata Using Micro-Aggregation”, *Journal of Official Statistics*, Vol.14, No.4, pp.449-461.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002) “Practical Data-oriented Microaggregation for Statistical Disclosure Control”, *IEEE Transactions on Knowledge and Data Engineering*, vol.14, no.1, pp.189-201.

Höhne (2003) “SAFE- A Method for Statistical Disclosure Limitation of Microdata”, Paper presented at Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, pp.1-3.
Ito, S., Isobe, S., Akiyama, H. (2008) “A Study on Effectiveness of Microaggregation as Disclosure Avoidance Methods: Based on National Survey of Family Income and Expenditure”, NSTAC Working Paper, No.10, pp.33-66 (in Japanese).
Ito, S. (2009) “On Microaggregation as Disclosure Avoidance Methods”, Journal of Economics, *Kumamoto Gakuen University*, Vol.15, No.3・4, pp.197-232 (in Japanese).
Makita, N., Ito, S., Horikawa, A., Goto, T., Yamaguchi, K. (2013) “Development of Synthetic Microdata for Educational Use in Japan”, Paper Presented at 2013 Joint IASE / IAOS Satellite Conference, Macau Tower, Macau, China, pp.1-9.