Creating Synthetic Microdata from Official Statistics: Random Number Generation in Consideration of Anscombe's Quartet

Kiyomi Shirakawa

Hitotsubashi University / National Statistics Center

Shinsuke Ito

Chuo University

Outline

1. Synthetic Microdata in Japan
2. Problems with Existing Synthetic Microdata
3. Correcting Existing Synthetic Microdata
4. Creating New Synthetic Microdata
5. Comparison between Various Sets of Synthetic Microdata
6. Conclusions and Future Outlook

1. Synthetic Microdata in Japan

Synthetic Microdata for educational use are available in Japan:

- ➢Generated using multidimensional statistical tables.
- ➢Based on the methodology of microaggregation

(Ito (2008), Ito and Takano (2011), Makita et al (2013))

- Created based on the original microdata from the 2004 'National Survey of Family Income and Expenditure'
- Synthetic microdata are not original microdata.

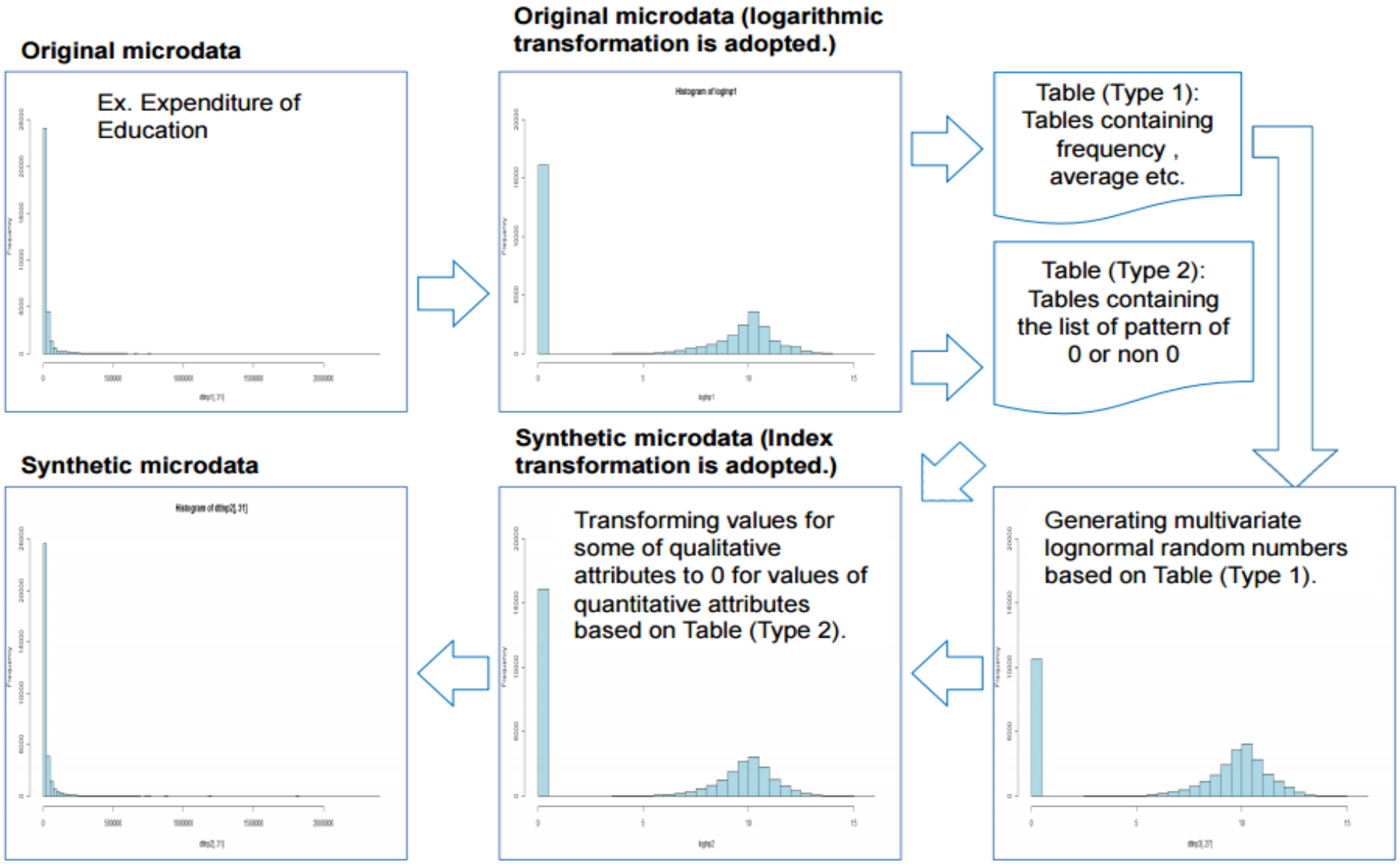
Legal Framework

New Statistics Act in Japan(April 2009)

- Enables the provision of Anonymized microdata (Article 36) and tailor-made tabulations (Article 34).
  - Allows a wider use of official microdata.
  - Allows use of official statistics in higher education and academic research.
  - However, permission process is required.

To provide an alternative to Anonymized microdata, the NSTAC has developed Synthetic microdata that can be accessed without a permission process.

Image of Frequency of Original and Synthetic Microdata



Source: Makita et al. (2013).

2. Problems with Existing Synthetic Microdata

(1) All variables are subjected to exponential transformation in units of cells in the result table.

Number of Earners	Structure of Dwelling	Frequency	Living Expenditure			Food		
			Mean	SD	C.V.	Mean	SD	C.V.
		4,132	302,492.8	148,598.9	0.491	71,009.0	25,089.5	0.353
One person	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
		4,201	346,195.7	215,911.7	0.624	78,209.1	25,288.1	0.323
Two persons	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		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

Too large

2. Problems with Existing Synthetic Microdata

(2) Correlation coefficients (numerical) between all variables are reproduced.

In the below table, several correlation coefficients are too small.

The reason is that correlation coefficients between uncorrelated variables are also reproduced.

	Living expenditure	Food	Housing
Living expenditure	1.00	0.50	0.28
Food	0.43	1.00	−0.03
Housing	0.28	−0.06	1.00

Too small

Top half: original data; bottom half: synthetic microdata

2. Problems with Existing Synthetic Microdata

(3) Qualitative attributes of groups having a frequency (size) of 1 or 2 are transformed to "Unknown" (V) or deleted.

The information loss when using this method is too large.

Furthermore, the variations within the groups are too large to merge qualitative attributes between different groups.

Individual Data

Number	Gender	Employment Status
1	1	1
2	1	1
3	1	1
4	1	3
5	1	4
6	1	4
:	:	:

→

Gender	Employment Status	N
1	1	3
1	3	1
1	4	2
:	:	:

→

Gender	Employment Status	N
1	1	3
1	V	3
:		:

→

Multidimensional Tables

Number	Gender	Employment Status
1	1	1
2	1	1
3	1	1
4	1	V
5	1	V
6	1	V
:	:	:

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.

3. Correcting Existing Synthetic Microdata

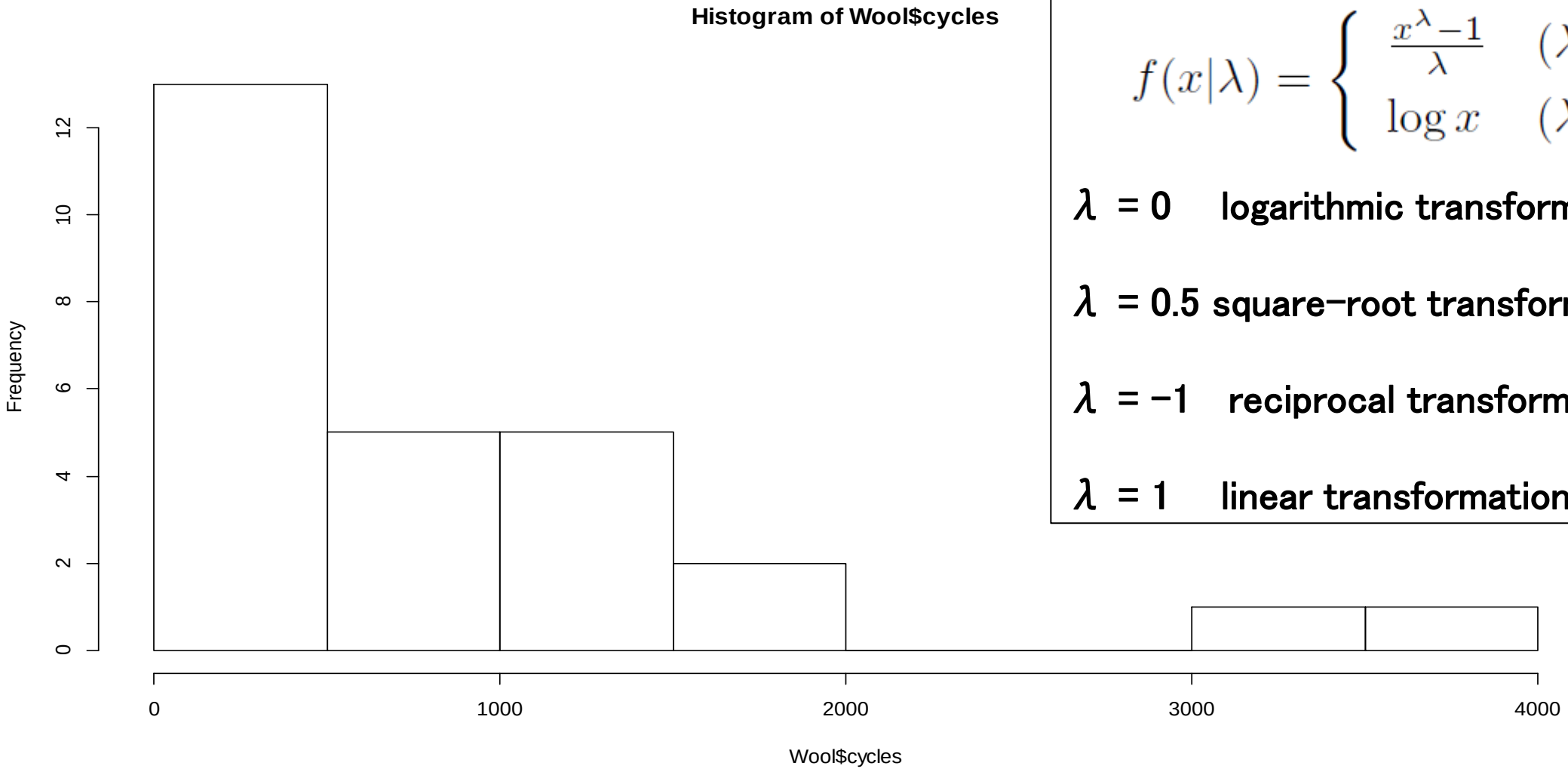
The following approaches can be used to correct the existing Synthetic microdata.

(1) Select the transformation method (logarithmic transformation, exponential transformation, square-root transformation, reciprocal transformation) based on the original distribution type (normal, bimodal, uniform, etc.).

(2) Detect non-correlations for each variable.

(3) Merge qualitative attributes in groups with a size of 1 or 2 into a group that has a minimum size of 3 in the upper hierarchical level.

Box–Cox Transformation



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

$\lambda = 0$ logarithmic transformation

$\lambda = 0.5$ square-root transformation

$\lambda = -1$ reciprocal transformation

$\lambda = 1$ linear transformation

4. Creating New Synthetic Microdata

In order to improve problems with existing Synthetic microdata, new synthetic microdata were created based on the following approaches.

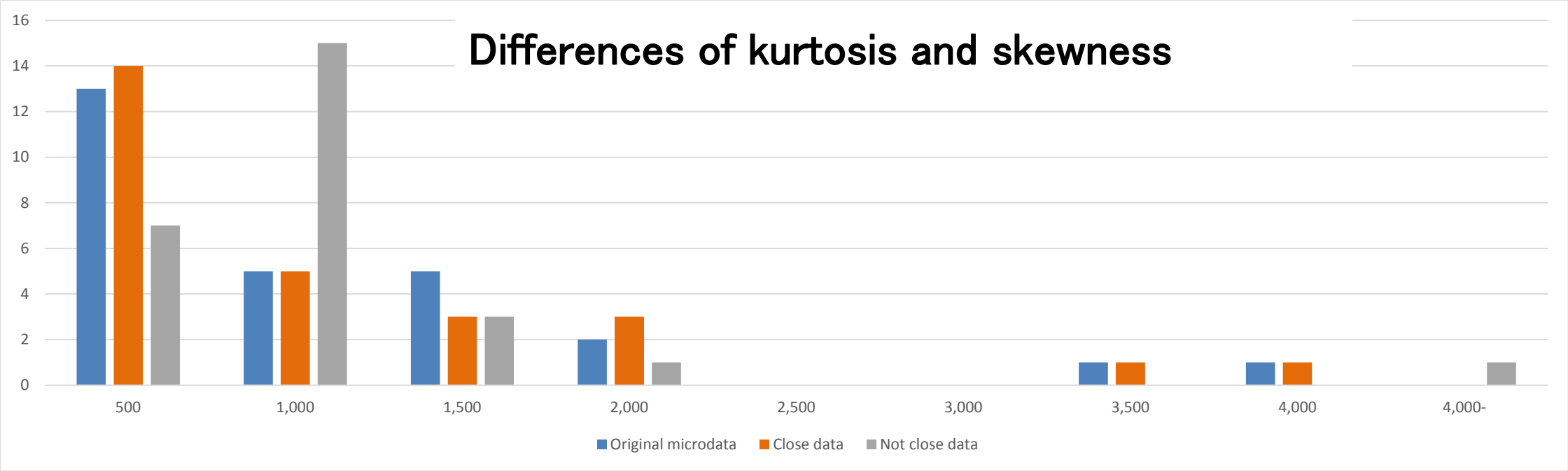
(1) Create microdata based on kurtosis and skewness

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

(3) Create microdata based on multivariate normal random numbers and exponential transformation

This process allows creating synthetic microdata with characteristics similar to those of the original microdata.

(1) Microdata created based on Kurtosis and Skewness



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
Standard deviation	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) Microdata created based on two Tabulation Tables (Basic Table and Details Table)

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
Standard deviation	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

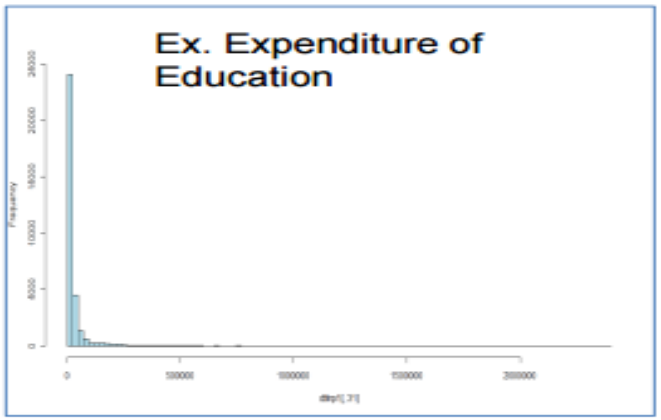
Details Table (means and standard deviations for creating synthetic microdata for multidimensional cross fields)

Groups	Living expenditure			Food		
	Frequency	Mean	Standard deviation	Frequency	Mean	Standard deviation
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

(3) Microdata created based on Multivariate Normal Random Numbers and Exponential Transformation

Original microdata

Ex. Expenditure of Education



Original microdata (logarithmic transformation is adopted.)

λ in the Box–Cox transformation is required in order to change the distribution type of the original data into a standard distribution.

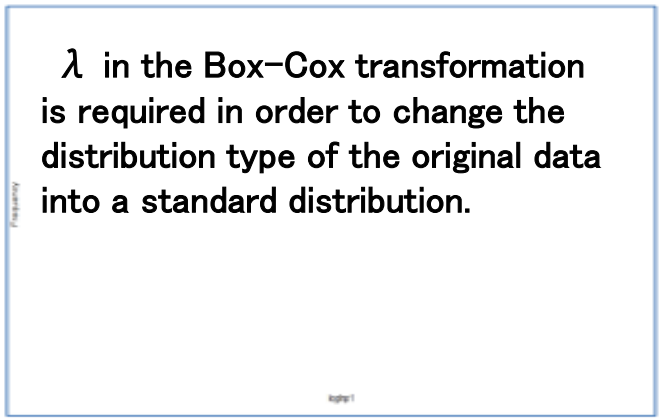


Table (Type 1): Tables containing frequency , average etc.

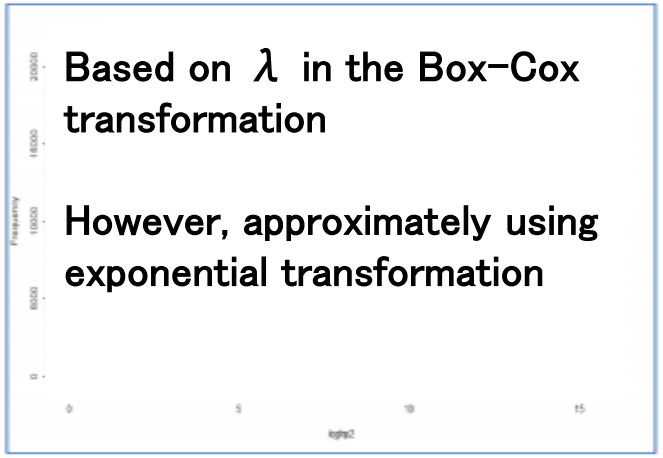
Table (Type 2): Tables containing the list of pattern of 0 or non 0

a random number that approximates the kurtosis and skewness of the original microdata was selected.

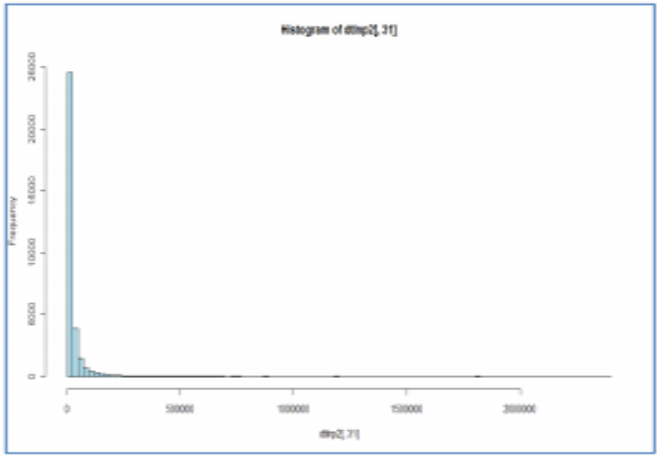
Synthetic microdata (Index transformation is adopted.)

Based on λ in the Box–Cox transformation

However, approximately using exponential transformation



Synthetic microdata



5. Comparison of Results

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

5. Comparison of Results

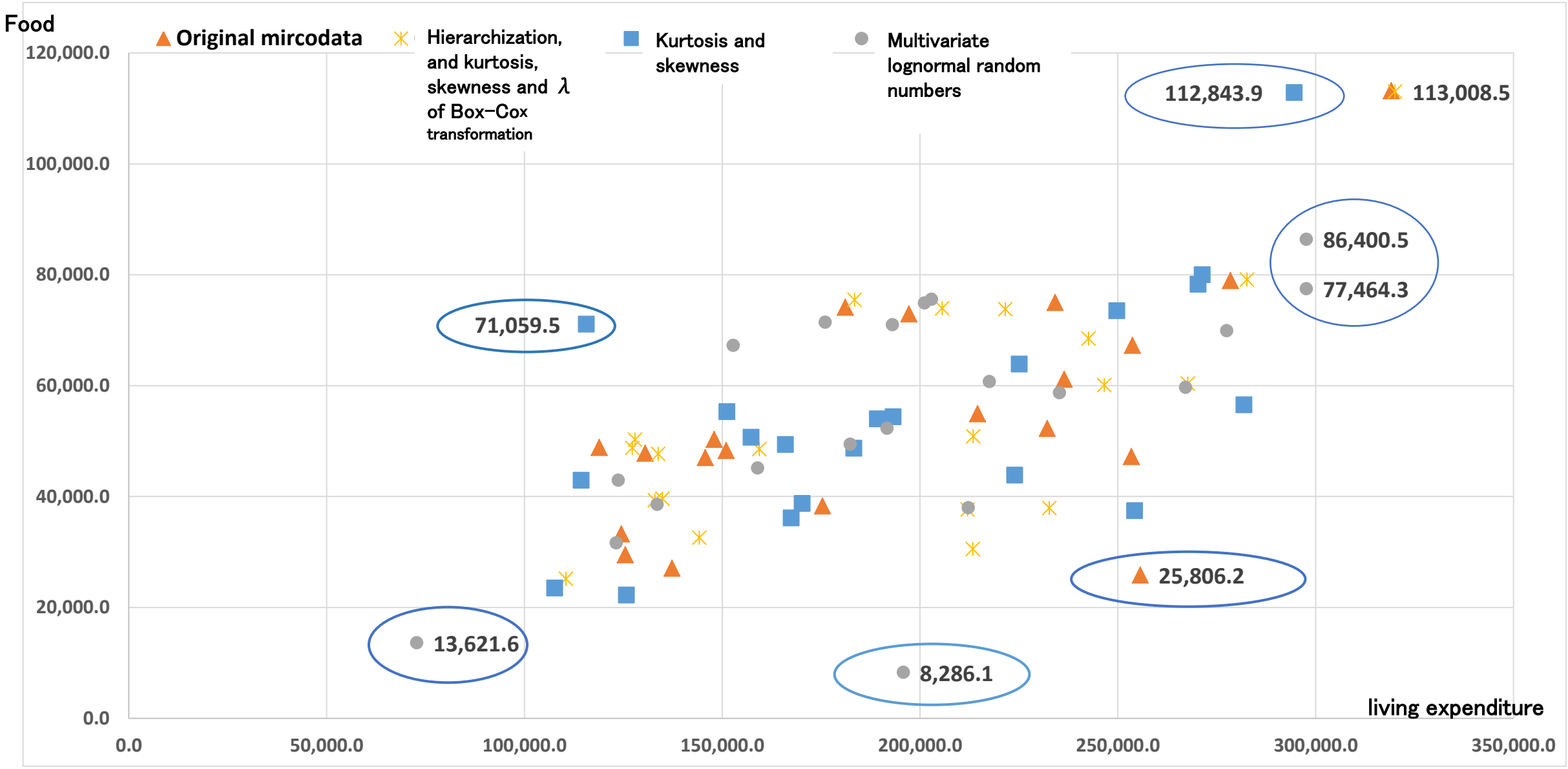
The most useful microdata from the indicators in the below table are in column number 2.

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
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
Standard deviation	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 value	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 value	118,895.1	25,806.2	110,487.8	25,143.0	107,684.0	22,188.5	72,773.7	8,286.1

Note that for reference, column number 4 is the same as the trial synthetic microdata method.

5. Comparison of Results

Scatter plots of living expenditure and food for each microdata



Example Result Table for New Synthetic Microdata

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					

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

6. Conclusions and Future Outlook

Conclusions

1. We suggested improvements to synthetic microdata created by the National Statistics Center for statistics education and training.
2. We created new synthetic microdata using several methods that adhere to this disclosure limitation method.
3. The results show that kurtosis, skewness, and Box–Cox transformation λ are useful for creating synthetic microdata in addition to frequency, mean, standard deviation, and correlation coefficient which have previously been used as indicators.

Next Steps

1. Decide the 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.
2. Expand this work to the creation and improvement of synthetic microdata from other surveys.

References

1. 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.
2. Defays, D. and Anwar, M.N.(1998) “Masking Microdata Using Micro–Aggregation”, *Journal of Official Statistics*, Vol.14, No.4, pp.449–461.
3. 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.
4. 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.
5. 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).
6. 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)
7. 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.