2013 Joint IASE / IAOS Satellite Conference, Macau Tower, Macau, China, 22nd-24th August, 2013

Development of Synthetic Microdata for Educational Use in Japan

Naoki Makita *National Statistics Center, Japan*

*Shinsuke Ito** *Meikai University/National Statistics Center, Japan*

Akiko Horikawa *National Statistics Center, Japan*

Takehiko Goto *National Statistics Center, Japan*

Kozo Yamaguchi Statistical Research and Training Institute Ministry of Internal Affairs and Communications, Japan

Outline

1. Summary

2. Background: Legal Framework

3. Synthetic Microdata for Educational Use

4. Creating Synthetic Microdata

5. Comparison: Original Microdata vs. Synthetic Microdata

6. Conclusion and Outlook

1. Summary

Statistics and Education

- Education and training of researchers who can conduct empirical analysis using microdata requires availability of microdata for educational use.
- However, due to usage restrictions and other issues, using Anonymized microdata for education and training is often not practical.
- Synthetic microdata offer unrestricted data that is freely available and therefore ideal for educational use.

Synthetic microdata can enhance and expand statistics education and training.

2. Background: Legal Framework

2.1. Japan's new Statistics Act (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.

The new Statistics Act expands the role of statistics in education and research in Japan.

2. Background: Legal Framework

2.2. The Role of the National Statistics Center

- The National Statistics Center (NSTAC) operates a data archive that provides Anonymized microdata and tailor-made tabulations using data collected by government offices and ministries.
- The NSTAC also cooperates with academic research organizations to promote this service.

The NSTAC plays a key role in the new framework for statistics education and research.

2. Background: Legal Framework

2.3. Procedures of Access to Anonymized Microdata

- Permission is required to access data.
- Burdensome application process.
- Strict conditions on data usage and storage.
- Costs involved.

➔ Not suitable for education and training.

- To provide an easier alternative, the NSTAC has developed Synthetic microdata that can be accessed without an permission process.
- These data do not contain Anoymized microdata.

Synthetic microdata solve many of the issues associated with Anonymized microdata.

3. Synthetic Microdata for Educational Use

3.1. Synthetic Microdata

- Generated using multidimensional statistical tables.
- Based on methodology of microaggregation (Ito(2008), Ito and Takano (2011))

Survey Data: Original microdata from the 2004 ‘National Survey of Family Income and Expenditure’

Synthetic microdata are not original microdata.

4. Creating Synthetic Microdata

(1) Selection of quantitative and qualitative attributes to be contained in Synthetic microdata.

(2) Sorting records with common values for qualitative attributes into groups with a minimum size of 3.

(3) Creation of tables in order to generate 1) multivariate lognormal random numbers and 2) records where values for some quantitative attributes are 0.

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

4. Creating Synthetic Microdata

(1) Selection of qualitative and quantitative attributes to be contained in Synthetic microdata.

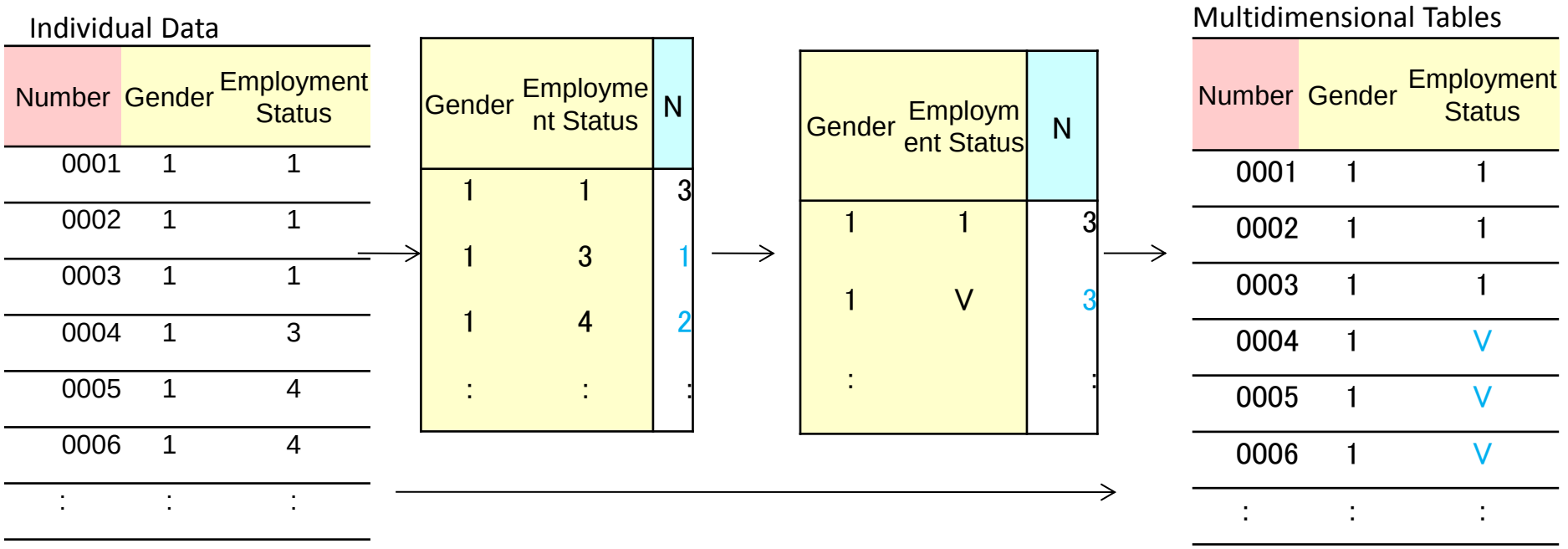
- Exploratory selection of qualitative attributes based on multidimensional statistical tables:

  14 qualitative attributes are selected based on frequency with which survey items are used by researchers, including gender, age and employment status.
- 184 quantitative attributes are also selected, including: yearly household income (‘yearly income’) and monthly household expenditures (‘living expenditure’).

Qualitative attributes that are most popular with researchers are used.

4. Creating Synthetic Microdata

(2) Processing records with common values for qualitative attributes into groups with a minimum size of 3.



For records with common values for qualitative attributes which refer to groups with size is 1 or 2, different values for some qualitative attributes are transformed to 'unknown' (V) in order to create groups with a minimum size of 3.

4. Creating Synthetic Microdata

(3) Creation of tables in order to generate 1) multivariate lognormal random numbers and 2) records where values for some quantitative attributes are 0.

Table (Type 1)

Tables which contain frequency, mean, variance and covariance of quantitative attributes not including 0. Base records are classified by qualitative attributes to generate multivariate lognormal random numbers.

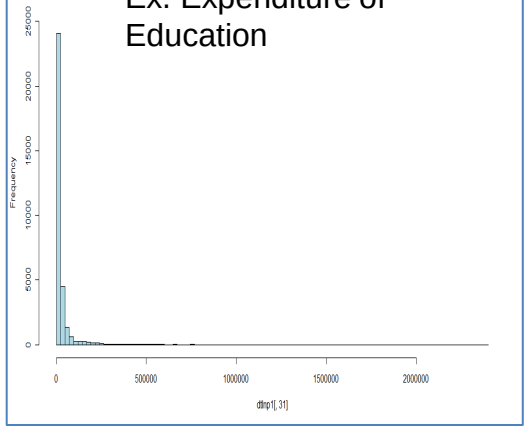
Table (Type 2)

Tables created based on whether values for quantitative attributes within a group are 0 or not 0, and according to this pattern the values for some quantitative attributes are transformed to 0.

Image of frequency of original microdata and Synthetic microdata

Original microdata

Ex. Expenditure of Education



Original microdata (logarithmic transformation is adopted.)

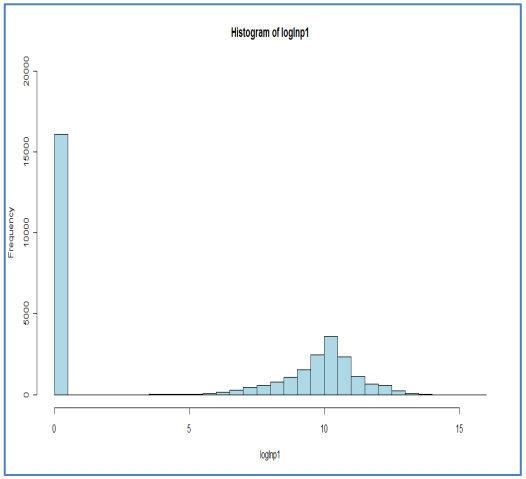
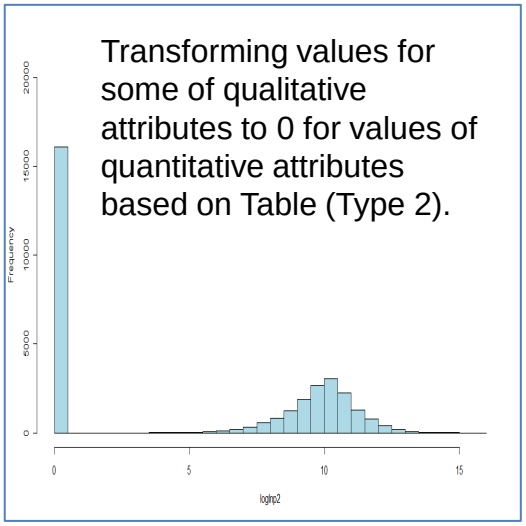


Table (Type 1): Tables containing frequency , average etc.

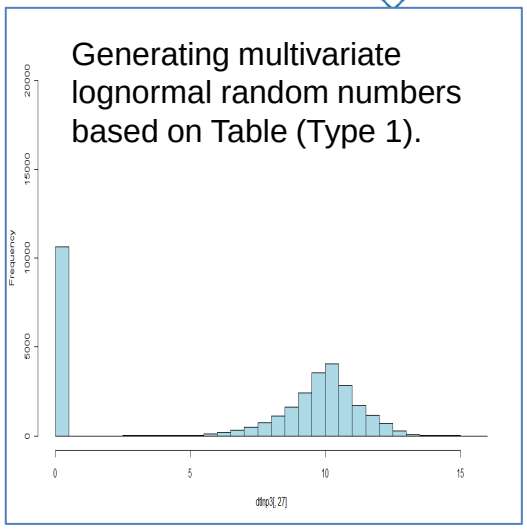
Table (Type 2): Tables containing the list of pattern of 0 or non 0

Synthetic microdata (Index transformation is adopted.)

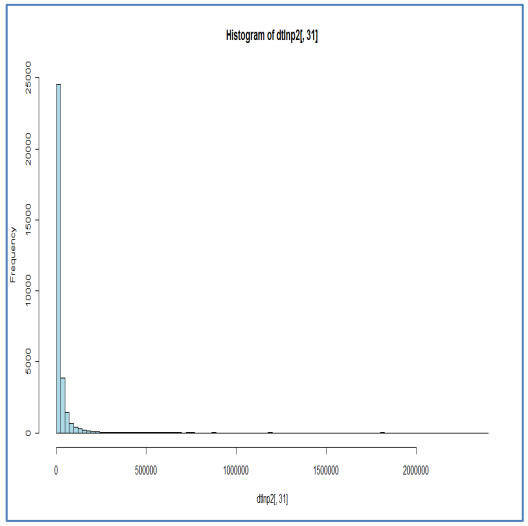
Transforming values for some of qualitative attributes to 0 for values of quantitative attributes based on Table (Type 2).



Generating multivariate lognormal random numbers based on Table (Type 1).



Synthetic microdata



5. Comparison: Original Microdata vs. Synthetic Microdata

5.1. Descriptive Statistics: Average Values

	Original Microdata	Synthetic Microdata	difference (%)
Yearly Income (ten thousand yen)	740	730	-0.01
Receipts (yen)	971,789	946,779	-0.03
Income (yen)	502,134	497,656	-0.01
Receipts other than Income (yen)	391,824	372,130	-0.05
Expenditure (yen)	415,809	403,747	-0.03
Living expenditure (yen)	339,199	328,140	-0.03
Non-living expenditure (yen)	76,610	75,607	-0.01

The averages of attributes contained in synthetic microdata are quite similar to those in original microdata.

5. Comparison: Original Microdata vs. Synthetic Microdata

5.2. Descriptive Statistics: Standard Deviation

	Original Microdata	Synthetic Microdata	difference (%)
Yearly Income (ten thousand yen)	358	338	-0.06
Receipts (yen)	541,291	473,481	-0.13
Income (yen)	280,696	261,558	-0.07
Receipts other than Income (yen)	353,922	263,446	-0.26
Expenditure (yen)	224,420	219,291	-0.02
Living expenditure (yen)	194,501	192,447	-0.01
Non-living expenditure (yen)	56,200	66,378	0.18

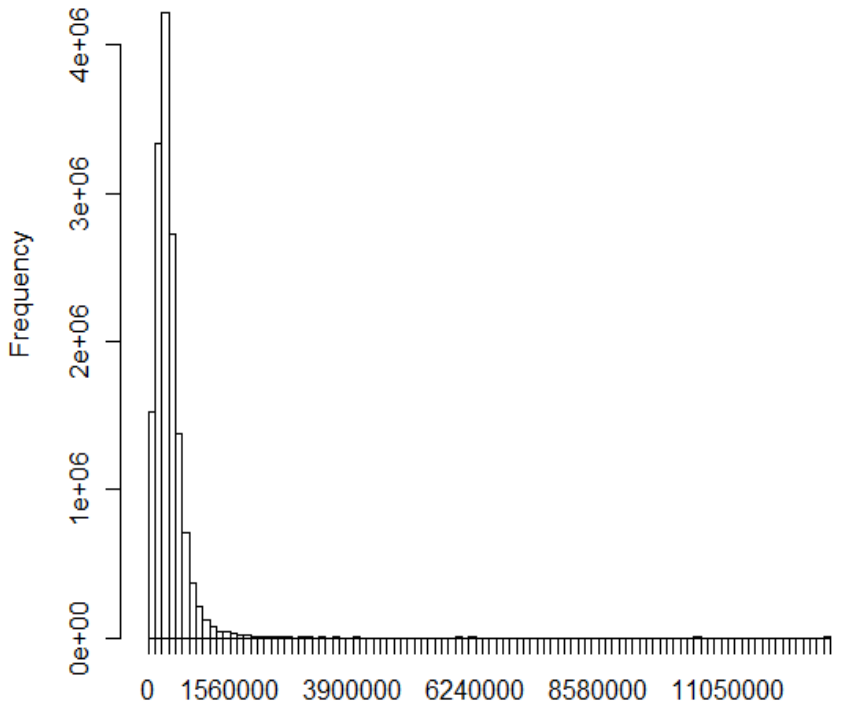
The standard deviation for Synthetic microdata is also similar to that for original microdata.

5. Comparison: Original Microdata vs. Synthetic Microdata

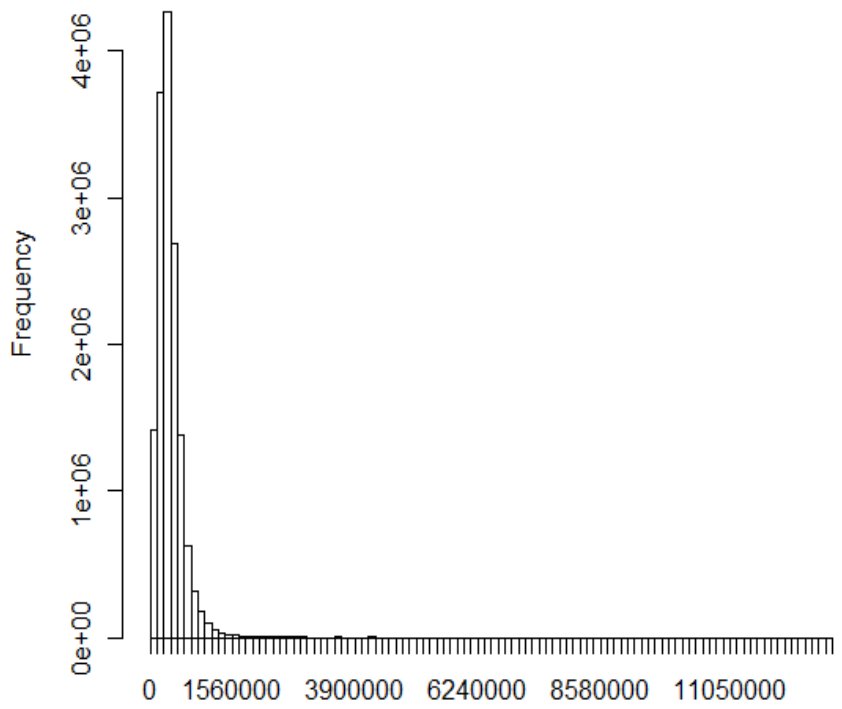
5.3. Histogram

Receipts other than Income

Mean 391,823.98
S.D. 353,922.37



Mean 372,130.47
S.D. 263,455.65



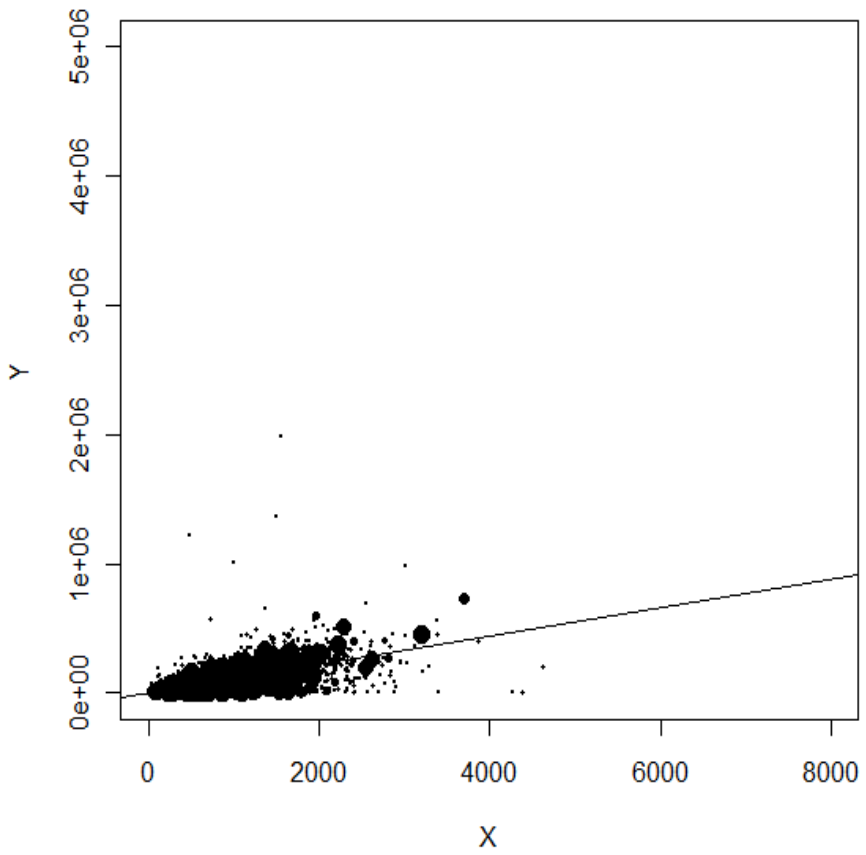
The histogram for 'Receipts other than Income' is similar for Synthetic microdata and original microdata.

5. Comparison: Original Microdata vs. Synthetic Microdata

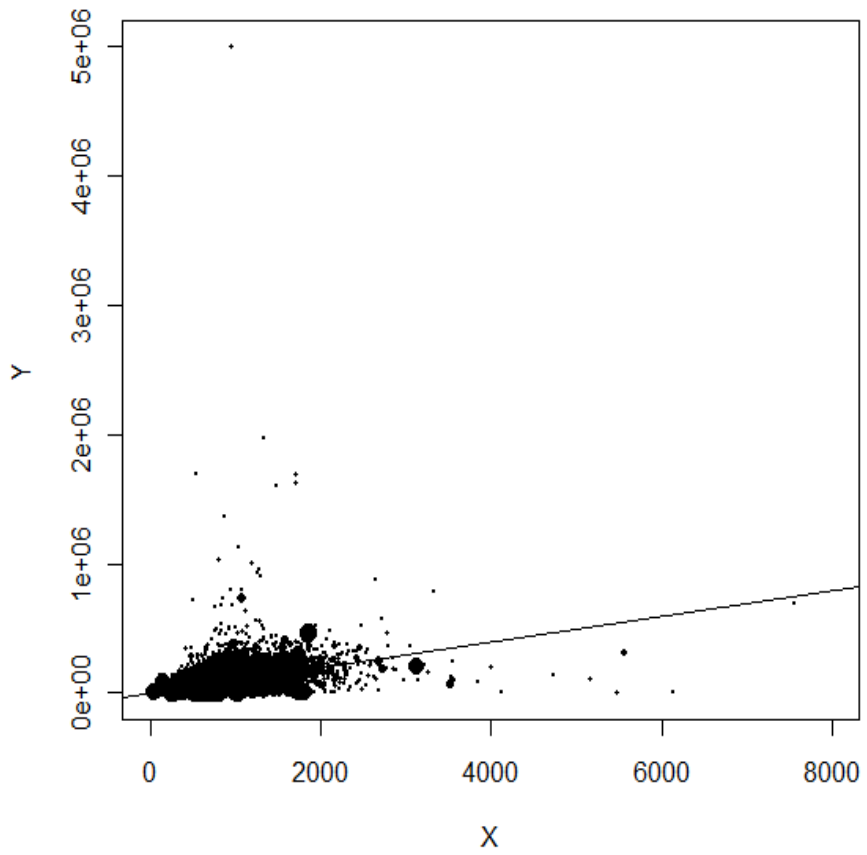
5.4. Scatter Diagram

Yearly Income (x) × Non-Living Expenditure (y)

Correlation Coefficient : 0.70



Correlation Coefficient : 0.50



For yearly income and non-living expenditure, there are more outliers for Synthetic microdata than for original microdata.

5. Comparison: Original Microdata vs. Synthetic Microdata

5.5. Correlation Coefficient

Original Microdata	Yearly Income	Receipts	Income	Receipts other than Income	Expenditure	Living expenditure	Non-living expenditure
Yearly Income	1.00						
Receipts	0.60	1.00					
Income	0.66	0.78	1.00				
Receipts other than Income	0.35	0.85	0.36	1.00			
Expenditure	0.60	0.73	0.56	0.63	1.00		
Living expenditure	0.49	0.66	0.45	0.61	0.97	1.00	
Non-living expenditure	0.70	0.63	0.70	0.38	0.62	0.43	1.00

5. Comparison: Original Microdata vs. Synthetic Microdata

5.5. Correlation Coefficient

Synthetic Microdata	Yearly Income	Receipts	Income	Receipts other than Income	Expenditure	Living expenditure	Non-living expenditure
Yearly Income	1.00						
Receipts	0.58	1.00					
Income	0.63	0.85	1.00				
Receipts other than Income	0.38	0.83	0.48	1.00			
Expenditure	0.52	0.71	0.59	0.64	1.00		
Living expenditure	0.42	0.63	0.49	0.60	0.96	1.00	
Non-living expenditure	0.50	0.50	0.52	0.35	0.53	0.26	1.00

For correlation matrix, the relationship between attributes in Synthetic microdata is maintained.

6. Conclusion and Outlook

- The new Statistics Act **expands the role of statistics** in education and research in Japan.
- Synthetic microdata are available **without restrictions and costs**.
- The process outlined in this research allows creating Synthetic microdata with characteristics similar to those of original microdata.
- Synthetic microdata have the potential to **enhance and expand** statistics education and training.