Creating of Synthetic Microdata Using Probabilistic Dummy Variables of High Dimensional Statistics based on Big data

30 October 2019

Kiyomi Shirakawa, Hitotsubashi University
Ken Nakamatsu, Takumi Information technology Corporation
Takayuki Ito, Hitotsubashi University
1. Existing method: Synthesis using random numbers

2. Trial: Synthesis using High Dimensional Statistics by Big Data
   - Synthesis using estimated Kurtosis and Skewness

Evaluation

Compare method # 1 and # 2: which give better results

Solving problems for practical use
1. Creating a Synthetic Microdata

- In Japan, we **cannot** create a synthetic microdata using an original microdata **directly**.
- We usually calculate fundamental statistics, such as frequency, mean, standard deviation and correlation coefficients, so on.

- **Existing method**: Synthesis using **random numbers**
1. Existing method

Image of Frequency of Original 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

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).

Source: Makita et al. (2013).
Creation of synthetic data using high-dimensional statistics, such as skewness and kurtosis

Tabular items: frequency, mean, SD, skewness, and kurtosis

Original data → tabulation → tables

Combination by freq. DB → extract candidate data → narrow

freq., Mean, SD, skewness, kurtosis

collate

store all of combination by frequency

because each value is integer, extract by each statistic approximation

GRG non-linear

using MS-EXCEL Solver

SD and skew. is the same, kurtosis is approximate

decide combi. of data
Frequency 20 of original and estimation (integer, real-number) values

Estimation of real-number: Calculated by MS-EXCEL solver

Source: Shirakawa et al. (2016).
However,

2. Normally, the skewness and kurtosis statistics are NOT tabulated in the official statistics results table.

Therefore,

3. Using Probabilistic Dummy Variables of High Dimensional Statistics, such as skewness and kurtosis Dummy Variable values are estimated by the big data.
Outline of the Big Data

Combination when N=20

● Requirement of combination assuming each variable is integer:
  \[ x_1 + x_2 + \cdots + x_{19} + x_{20} = 100 \]
  \[ x_{20} \leq x_{19} \leq \cdots \leq x_2 \leq x_1 \]

● Range of each variable
  N-th maximum and minimum value

|       | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | X_7 | X_8 | X_9 | X_{10} | X_{11} | X_{12} | X_{13} | X_{14} | X_{15} | X_{16} | X_{17} | X_{18} | X_{19} | X_{20} |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Max.  | 100 | 50  | 33  | 25  | 20  | 16  | 14  | 12  | 11  | 10     | 9      | 8      | 7      | 7      | 6      | 6      | 5      | 5      | 5      | 5      |
| Mini. | 5   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0   | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      | 0      |

● Number of Combinations: 97,132,873 patterns
## Number of combination by frequency

<table>
<thead>
<tr>
<th>frequency</th>
<th># of combi.</th>
<th>$X_f$ (largest variable in combination)</th>
<th>Max.</th>
<th>Mini.</th>
<th>mode</th>
<th># of combination of mode</th>
<th>ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>97,132,873</td>
<td></td>
<td>100</td>
<td>5</td>
<td>20</td>
<td>5,498,387</td>
<td>5.7</td>
</tr>
<tr>
<td>19</td>
<td>86,658,411</td>
<td></td>
<td>100</td>
<td>6</td>
<td>20</td>
<td>4,865,168</td>
<td>5.6</td>
</tr>
<tr>
<td>18</td>
<td>75,772,412</td>
<td></td>
<td>100</td>
<td>6</td>
<td>21</td>
<td>4,214,238</td>
<td>5.6</td>
</tr>
<tr>
<td>17</td>
<td>64,684,584</td>
<td></td>
<td>100</td>
<td>6</td>
<td>21</td>
<td>3,565,531</td>
<td>5.5</td>
</tr>
<tr>
<td>16</td>
<td>53,662,038</td>
<td></td>
<td>100</td>
<td>7</td>
<td>22</td>
<td>2,914,553</td>
<td>5.4</td>
</tr>
<tr>
<td>15</td>
<td>43,018,955</td>
<td></td>
<td>100</td>
<td>7</td>
<td>22</td>
<td>2,307,803</td>
<td>5.4</td>
</tr>
<tr>
<td>14</td>
<td>33,097,743</td>
<td></td>
<td>100</td>
<td>8</td>
<td>23</td>
<td>1,743,387</td>
<td>5.3</td>
</tr>
<tr>
<td>13</td>
<td>24,234,058</td>
<td></td>
<td>100</td>
<td>8</td>
<td>23</td>
<td>1,248,206</td>
<td>5.2</td>
</tr>
<tr>
<td>12</td>
<td>16,713,148</td>
<td></td>
<td>100</td>
<td>9</td>
<td>24</td>
<td>840,657</td>
<td>5.0</td>
</tr>
<tr>
<td>11</td>
<td>10,718,685</td>
<td></td>
<td>100</td>
<td>10</td>
<td>25</td>
<td>523,214</td>
<td>4.9</td>
</tr>
<tr>
<td>10</td>
<td>6,292,069</td>
<td></td>
<td>100</td>
<td>10</td>
<td>27</td>
<td>296,777</td>
<td>4.7</td>
</tr>
<tr>
<td>9</td>
<td>3,314,203</td>
<td></td>
<td>100</td>
<td>12</td>
<td>28</td>
<td>149,938</td>
<td>4.5</td>
</tr>
<tr>
<td>8</td>
<td>1,527,675</td>
<td></td>
<td>100</td>
<td>12</td>
<td>30</td>
<td>65,728</td>
<td>4.3</td>
</tr>
<tr>
<td>7</td>
<td>596,763</td>
<td></td>
<td>100</td>
<td>15</td>
<td>32</td>
<td>24,135</td>
<td>4.0</td>
</tr>
<tr>
<td>6</td>
<td>189,509</td>
<td></td>
<td>100</td>
<td>17</td>
<td>36</td>
<td>7,119</td>
<td>3.8</td>
</tr>
<tr>
<td>5</td>
<td>46,262</td>
<td></td>
<td>100</td>
<td>20</td>
<td>40</td>
<td>1,592</td>
<td>3.4</td>
</tr>
<tr>
<td>4</td>
<td>8,037</td>
<td></td>
<td>100</td>
<td>25</td>
<td>46</td>
<td>251</td>
<td>3.1</td>
</tr>
<tr>
<td>3</td>
<td>884</td>
<td></td>
<td>100</td>
<td>34</td>
<td>50</td>
<td>26</td>
<td>2.9</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
<td></td>
<td>100</td>
<td>50</td>
<td>—</td>
<td>1</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Unsafe combinations by \( p\% \) rule ( freq. \( N=20 \))

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>freq.</th>
<th>( X_1 )</th>
<th>freq.</th>
<th>( X_1 )</th>
<th>freq.</th>
<th>( X_1 )</th>
<th>freq.</th>
<th>( X_1 )</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>89</td>
<td>19</td>
<td>78</td>
<td>195</td>
<td>67</td>
<td>373</td>
<td>56</td>
<td>195</td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td>88</td>
<td>30</td>
<td>77</td>
<td>195</td>
<td>66</td>
<td>373</td>
<td>55</td>
<td>139</td>
</tr>
<tr>
<td>98</td>
<td>2</td>
<td>87</td>
<td>30</td>
<td>76</td>
<td>272</td>
<td>65</td>
<td>272</td>
<td>54</td>
<td>139</td>
</tr>
<tr>
<td>97</td>
<td>2</td>
<td>86</td>
<td>45</td>
<td>75</td>
<td>272</td>
<td>64</td>
<td>272</td>
<td>53</td>
<td>139</td>
</tr>
<tr>
<td>96</td>
<td>4</td>
<td>85</td>
<td>45</td>
<td>74</td>
<td>373</td>
<td>63</td>
<td>272</td>
<td>52</td>
<td>139</td>
</tr>
<tr>
<td>95</td>
<td>4</td>
<td>84</td>
<td>67</td>
<td>73</td>
<td>373</td>
<td>62</td>
<td>272</td>
<td>51</td>
<td>139</td>
</tr>
<tr>
<td>94</td>
<td>7</td>
<td>83</td>
<td>67</td>
<td>72</td>
<td>508</td>
<td>61</td>
<td>272</td>
<td>50</td>
<td>97</td>
</tr>
<tr>
<td>93</td>
<td>7</td>
<td>82</td>
<td>97</td>
<td>71</td>
<td>508</td>
<td>60</td>
<td>195</td>
<td>49</td>
<td>95</td>
</tr>
<tr>
<td>92</td>
<td>12</td>
<td>81</td>
<td>97</td>
<td>70</td>
<td>373</td>
<td>59</td>
<td>195</td>
<td>48</td>
<td>90</td>
</tr>
<tr>
<td>91</td>
<td>12</td>
<td>80</td>
<td>139</td>
<td>69</td>
<td>373</td>
<td>58</td>
<td>195</td>
<td>46</td>
<td>52</td>
</tr>
<tr>
<td>90</td>
<td>19</td>
<td>79</td>
<td>139</td>
<td>68</td>
<td>373</td>
<td>57</td>
<td>195</td>
<td>47</td>
<td>78</td>
</tr>
</tbody>
</table>

Total 8,849

It’s only 8,849 sets in 97,132,873 sets (0.01%)
Using Original data

Toyo Keizai Corporate Segment Data for 5 years
Total number of records approximately 34,000
4 variables:
  Sales, Profits, Assets, number of employees

2 attributes
  1. Number of industry codes: 33 categories
  2. Grouping: Frequency 10 for each industry
**Utilization of Big Data**

<table>
<thead>
<tr>
<th>Patterns n=10, sum=100</th>
<th>Standard Deviation</th>
<th>Standard Deviation ratio of maximum value</th>
<th>kurtosis</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 0 ... 0</td>
<td>31.623</td>
<td>1</td>
<td>10</td>
<td>3.162</td>
</tr>
<tr>
<td>99 1 ... 0</td>
<td>31.273</td>
<td>0.989</td>
<td>9.997</td>
<td>3.162</td>
</tr>
<tr>
<td>... ... ...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>11 10 ... 9</td>
<td>0.471</td>
<td>0.015</td>
<td>0</td>
<td>4.500</td>
</tr>
<tr>
<td>10 10 ... 10</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Standard Deviation**

1. Calculation of maximum value ratio
   Value of each pattern / maximum value (31.623)

2. Extraction of target pattern
   Target pattern ratio : ± 0.005
## Utilization of Big Data

<table>
<thead>
<tr>
<th>Patterns n=10, sum=100</th>
<th>Standard Deviation</th>
<th>Standard Deviation ratio of maximum value</th>
<th>kurtosis</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 0 ... 0</td>
<td>31.623</td>
<td>1</td>
<td>10</td>
<td>3.162</td>
</tr>
<tr>
<td>99 1 ... 0</td>
<td>31.273</td>
<td><strong>0.989</strong></td>
<td>9.997</td>
<td>3.162</td>
</tr>
<tr>
<td>... ... ...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>11 10 ... 9</td>
<td>0.471</td>
<td>0.015</td>
<td>0</td>
<td>4.500</td>
</tr>
<tr>
<td>10 10 ... 10</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Standard Deviation

1. Calculation of maximum value ratio
   - Value of each pattern / maximum value (31.623)
   - $31.273 / 31.623 = 0.989$

2. Extraction of target pattern
   - Target pattern ratio : $\pm 0.005$
# Utilization of Big Data

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Standard Deviation</th>
<th>Standard Deviation ratio of maximum value</th>
<th>Kurtosis</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=10, sum=100</td>
<td>100</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>10</td>
<td>...</td>
<td>9</td>
</tr>
</tbody>
</table>

### Standard Deviation

1. For example
   - mean: 5,000, SD: 200, N=10
   
   $\frac{200}{5,000} \times 10 \div 31.623 = 0.013$

2. Extraction range ($\pm 0.005$)
   - Between **0.008** and **0.018**
### Estimation of kurtosis and skewness by Big Data

#### Example

<table>
<thead>
<tr>
<th>Example</th>
<th>mean</th>
<th>SD</th>
<th>SD ratio</th>
<th>Kurt</th>
<th>Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37,747.7</td>
<td>38,797.5</td>
<td>0.325</td>
<td>-0.40</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>30,770.5</td>
<td>42,899.9</td>
<td>0.441</td>
<td>4.35</td>
<td>2.09</td>
</tr>
</tbody>
</table>

#### Standard Deviation Extraction range

**Example 1**
- Between 0.320 and 0.330
- Number of Extraction 277,829 rcds

**Example 2**
- Between 0.436 and 0.446
- Number of Extraction 98,717 rcds
Prioritize skewness of concentrated values

- The range of skewness is limited.
- The reference value may be either the mode value or the median value.

Original 0.97
Extraction range: 0.47 - 1.47

Original 2.09
Extraction range: 1.59 - 2.59
Extraction by kurtosis

Number of records after extraction by skewness

- **Original**: 178,271 rcds (277,829)
- **Median**: 0.074
- **Freq.**

- **Original**: 60,186 rcds (98,717)
- **Median**: 2.661
- **Freq.**

In kurtosis, the mode value and median value differ.
### Extraction results

Extracting patterns that actually exist in Big Data from estimated values

#### Ex1

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Estimated Value</th>
<th>Extracting Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurt</td>
<td>-0.40</td>
<td>0.2396</td>
<td>0.2426</td>
</tr>
<tr>
<td>Skew</td>
<td>0.97</td>
<td>1.1187</td>
<td>1.1187</td>
</tr>
</tbody>
</table>

#### Ex2

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Estimated Value</th>
<th>Extracting Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurt</td>
<td>4.35</td>
<td>2.6617</td>
<td>2.6757</td>
</tr>
<tr>
<td>Skew</td>
<td>2.09</td>
<td>1.7762</td>
<td>1.7767</td>
</tr>
</tbody>
</table>

**Range**
- Ex1: $-0.426 - 0.574$
- Ex2: $2.161 - 3.161$

**Medians**
- Ex1: 0.074
- Ex2: 2.661

**Median**
- Ex1: 0.074
- Ex2: 2.661
Execution environment

• Execution time for extracting patterns from DB is a bottleneck
  4 variables : about 34,000 records
  \[ 4 \times 34,000 / 10 \text{ (N per group)} = 13,600 \text{ times} \]

• DB uses sqlite3 in memory
  \( N = 10 \text{ pattern, about 6.3 million records} \)
  Processing time: 1 second / query level
  Total processing time : around 5-6 hours

• PC used for calculation
  OS : win10
  CPU : Intel Xeon X5460 (4Core)
  Memory : 32GB
Synthesis result (N=10): distribution example 1

Industry: Agriculture, Forestry and Fisheries Group 1
Synthesis result (N=10): distribution example 2

Industry: Construction Group: 3

Original data

Big data

Random data
Comparison result 1: Absolute value difference

- **Sales**: 64.1% (Big Data) vs. 68.0% (Random)
- **Profits**: 64.8% (Big Data) vs. 68.0% (Random)
- **Assets**: 64.8% (Big Data) vs. 57.5% (Random)
- **Number of employees**: 57.5% (Big Data) vs. 57.5% (Random)
Comparison result 2: Excessive difference by evaluation method

• Overall evaluation of the two methods
  ➢ Evaluation criterion for comparison

• Numerical gap between groups
  \[
  \left( \frac{(\text{Synthesis data} - \text{original data})}{\text{original data}} \right)^2
  \]

• Small values have large errors
  If so, exclude from evaluation

• Example of error for each record when the absolute difference is the same
  For example,
  
  Original data 100,000 synthesis 101,000 \(\rightarrow\) 0.0001
  Original data 1 synthesis 1,001 \(\rightarrow\) 1,000,000

• Criterion for missing values
  \[
  \left( \frac{(\text{Original data} - \text{synthetic data})}{\text{original data}} \right)^2 > 100
  \]
Missing rate (%) (deletion rate due to large difference)

### Sales
- **Big data**
  - Missing rate: 1.737
- **Random data**
  - Missing rate: 4.287

### Profits
- **Big data**
  - Missing rate: 4.131
- **Random data**
  - Missing rate: 7.167

### Assets
- **Big data**
  - Missing rate: 2.587
- **Random data**
  - Missing rate: 5.734

### Number of employees
- **Big data**
  - Missing rate: 2.077
- **Random data**
  - Missing rate: 4.580
Comparison result 3: Evaluation of kurtosis and skewness

- Evaluation by synthesis unit
- Evaluation criteria: approximation to original data
- Calculate kurtosis and skewness for each group
- Comparison by mean square error

$$\sqrt{\frac{1}{n}\sum (\text{Synthetic data} - \text{Original data})^2}$$
Mean square error (kurtosis)

- **Sales**
  - Random data: 7.445
  - **Big data**: 2.540

- **Assets**
  - Random data: 5.911
  - **Big data**: 2.649

- **Profits**
  - Random data: 7.096
  - **Big data**: 2.656

- **Number of employees**
  - Random data: 5.967
  - **Big data**: 2.596
Mean square error (Skewness)

### Sales
- **Big data**: Mean square error = 0.617
- **Random data**: Mean square error = 1.780

### Profits
- **Big data**: Mean square error = 0.642
- **Random data**: Mean square error = 1.976

### Assets
- **Big data**: Mean square error = 0.630
- **Random data**: Mean square error = 1.453

### Number of employees
- **Big data**: Mean square error = 0.625
- **Random data**: Mean square error = 1.474
Conclusion

• Large number of data
  Both methods have no difference because \( N \) is large.

• Small number of data
  The method for estimating kurtosis and skewness is closer to the original data value than the method using random numbers.

• Execution time
  Extraction process takes time by using Big data. Reduces extraction time by executing each frequency unit.
Future works

1. Extracting pattern candidates from standard deviation
   - Reduce Execution time
   - Effective extraction method from big data

2. Method for estimating kurtosis skewness
   - Exploration of the other indicators
     - Indicator combined with median and mode value

3. Correlation and sorting
   - Keep correlation of each variable
   - Sorting numbers in a pattern

4. Method for evaluating synthetic data
   - Difference in overall distribution and individual values
Thank you for your attention

kshirakawa@ier.hit-u.ac.jp