Carnegie Mellon University
The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim (Department of Statistics)

December 14, 2015
This work was partially supported by a NSF-Census Research Network grant, NSF SES 1130706. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.
What is data swapping?

- An algorithm to protect respondents in a dataset from risk of identification
- Applied by Census Bureau to Decennial Census and American Community Survey data before public release
Goal

- Ideally, study the Census Bureau’s own data swapping algorithm
- Cannot due to necessary secrecy, so create a similar algorithm and study it instead
High-level overview

- Calculate chi-squared value for each tract’s table, e.g. of age by marital status for census tracts #1–500
- Apply data swapping algorithm hundreds of times for each combination of swap rate and table
- Compare the resulting distribution of chi-squared values to the original value
- How well can we characterize the effect of data swapping on contingency tables?
Example of data swapping

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Sex</th>
<th>Income</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvey Dent</td>
<td>26-30</td>
<td>M</td>
<td>$150,000</td>
<td>South Gotham</td>
</tr>
<tr>
<td>Bruce Wayne</td>
<td>26-30</td>
<td>M</td>
<td>$102 Million</td>
<td>North Gotham</td>
</tr>
</tbody>
</table>
Example of data swapping

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Income</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvey Dent</td>
<td>26-30</td>
<td>M</td>
<td>$150,000</td>
</tr>
<tr>
<td>Bruce Wayne</td>
<td>26-30</td>
<td>M</td>
<td>$102 Million</td>
</tr>
</tbody>
</table>

↓

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Income</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvey Dent</td>
<td>26-30</td>
<td>M</td>
<td>$150,000</td>
</tr>
<tr>
<td>Bruce Wayne</td>
<td>26-30</td>
<td>M</td>
<td>$102 Million</td>
</tr>
</tbody>
</table>
Effect on contingency tables

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th></th>
<th>South</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Young</td>
<td>Old</td>
<td>Young</td>
<td>Old</td>
</tr>
<tr>
<td>Normal</td>
<td>500</td>
<td>200</td>
<td>100</td>
<td>400</td>
</tr>
<tr>
<td>Millionaire</td>
<td>1</td>
<td>20</td>
<td>4</td>
<td>30</td>
</tr>
</tbody>
</table>
Effect on contingency tables

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Young</td>
<td>Old</td>
</tr>
<tr>
<td>Normal</td>
<td>500</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Millionaire</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

↓

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Young</td>
<td>Old</td>
</tr>
<tr>
<td>Normal</td>
<td>501</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Millionaire</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pittsburgh, Allegheny County, PA

= one Public Use Microdata Area, or PUMA
= nine PUMAs
Simulated data swapping on two different datasets

1. Public Use Microdata Sample (PUMS)
   - Sub-sample of the collected American Community Survey (ACS) data
   - Collected between 2007–2011 for Allegheny County, PA
   - Lacks real tract information, only has PUMAs
   - We add artificial tract information

2. Full unswapped ACS data
   - Contains accurate, unswapped geography
   - Available to researchers on a need-to-know basis
Using $V$ instead of $\chi^2$

Thousands of tables with different $\chi^2$ values, so we need a way of normalizing them:

$$V = \sqrt{\frac{\chi^2/N}{\min(k-1, r-1)}}.$$

$N$ is the total number of people in a table with $r$ rows and $k$ columns.
$V$ ranges from $0$ (total independence; no association) to $1$ (the two variables are equal).
Stages of *non-targeted* data swapping

1. **Matching**: Randomly pair a fixed proportion of individuals, given they match on a set of attributes and their households match in size

2. **Swapping**: Swap the addresses of all of the individuals in the households
Stages of *targeted* data swapping

1. **Targeting:** Rank the individuals in data by disclosure risk
2. **Matching:** Randomly pair a fixed proportion of them, given they match on a set of attributes and their households match in size
3. **Swapping:** Swap the addresses of all of the individuals in the households
### Contingency table of age by marital status

<table>
<thead>
<tr>
<th>Age</th>
<th>Married</th>
<th>Widowed</th>
<th>Divorced</th>
<th>Separated</th>
<th>Never Married</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 16</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>194</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>269</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>188</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>153</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>111</td>
</tr>
<tr>
<td>23</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>24</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>93</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>≥ 94</td>
<td>3</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Effectiveness of non-targeted swapping

The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim
(Department of Statistics)

Introduction
Data
Methodology
Results
Conclusion

Swap Rate
Probability of Swapping

0.05  0.06  0.07  0.08  0.09  0.10  0.11  0.12  0.13  0.14  0.15

0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7

Swap Rate
Probability of Swapping

0.05  0.06  0.07  0.08  0.09  0.10  0.11  0.12  0.13  0.14  0.15
The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim (Department of Statistics)

Introduction

Data

Methodology

Results

Conclusion

Effectiveness of targeted swapping

![Box plot showing the effect of targeted swapping on contingency table analyses.](image)
The following plots were generated from simulated swaps on the public-use data.

Recall that these data do not have census geography at a sufficiently detailed level.
The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim (Department of Statistics)

Introduction

Data

Methodology

Results

Conclusion

Trajectory of Cramér’s V for non-targeted swapping

Blue line = original Cramér’s V value
The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim (Department of Statistics)

Introduction

Methodology

Results

Conclusion

Trajectory of Cramér’s $V$ for non-targeted swapping

Blue line = original Cramér’s $V$ value
Generality of this effect

- The Census data are very complicated and even randomized swapping has to keep household sizes constant and match on variables
- Do we observe this effect in the simplest scenario?
Simple data, simple swap

- Generated data with age, income, and a tract variable
- The incomes are a linear function of age, but the slope and amount of added noise are dependent on tract
- Some have strong association, some do not
- Data swapping algorithm only swaps geography without trying to match characteristics, and without targeting at-risk individuals
- Following plot was generated with this simple data
The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim (Department of Statistics)

Introduction

Data

Methodology

Results

Conclusion

Cramer's V after data swapping at increasing rates

Green line = average Cramér's V value across tracts
Public Use Microdata Sample, targeted swap

- With the non-targeted swap, we could say that $\chi^2$ and $V$ were pushed to a central value.
- What if we try swapping at-risk individuals with a targeted swapping algorithm?
- Do we still observe behavior that is easy to explain, and therefore easy to “undo”?
The Effect of Data Swapping on Contingency Table Analyses

Nicolás Kim (Department of Statistics)

Introduction
Data Methodology
Results
Conclusion

Trajectory of Cramér’s $V$ for targeted swapping

![Graph showing the trajectory of Cramér’s $V$ for targeted swapping. The graph plots Cramér’s $V$ against the swap rate, with a peak around a swap rate of 0.08, declining thereafter.]
Trajectory of Cramér’s $V$ for targeted swapping
Even for relatively low rates of swapping, like 0.05, we cannot say what effect targeted swapping has on the data. Perhaps this was an artifact of the artificial tract assignments we created? We ran the targeted swap simulations for 1600 swaps per swap rate for each half-integer swap rate between 5% and 15% (0.05 to 0.15).
Effect still present in real, unswapped ACS data

CV for MAR&RCGP, Tract 270100, n = 1600
Effect still present in real, unswapped ACS data
Future work

- $\chi^2$ and $V$ are natural statistics that capture the strength of association between two categorical variables, but other statistical procedures should be evaluated.

- A different targeting scheme might preserve the intuitive behavior of the non-targeted swap, but probably not for all tables and statistical procedures (on the other hand, a simple targeting criterium does not preserve $\chi^2$ statistics).
Conclusions

- **Data users**: be on the lookout for discrepancies in data that may be explained by data swapping; be aware of potential attenuations of joint distributions, although we cannot say in which way they are attenuated.

- **Census agencies**: computers are improving much faster than the population is growing: evaluate as many statistical procedures on as many forms of the data as possible, pre- and post-swapping. In addition, some Census Bureau researchers are now looking into the effect of swapping on various statistical measures (Lemons, Dajani, You 2015); in the future, these analyses should happen for the true data swapping algorithm.
Thank you!

Questions?