FCSM Federal Committee on Statistical Methodology

Confidentiality and Data Access Committee:
2017 Workshop on New Advances in Disclosure Limitation
September 27, 2017

Workshop Report

Executive Summary

On September 27, 2017, the Federal Committee on Statistical Methodology's Committee on Data Access and Confidentiality held a day-long workshop at the Bureau of Labor Statistics Conference Center to discuss the current state of statistical disclosure limitation. Approximately 80 people attended, most from the U.S. Government, with the US Census Bureau receiving the most representation.

Background

Today disclosure limitation within the Federal Statistical Community is largely based on the principles outlined in Statistical Policy Working Paper 22, *Report on Statistical Disclosure Limitation Methodology*, last revised in 2005. One of the key assumptions for data releases that Working Paper 22 makes is that the statistical agency is the primary holder of information that is used for the compilation of statistical products. As a result, the release methodologies described in Working Paper 22 largely depend upon belief that techniques such as top-coding, aggregation, and cell suppression are sufficient for mathematically protecting respondent data. As for microdata, few approaches are provided that strongly limit disclosure risk other than simply not releasing microdata, recoding the data to eliminate sample uniques, and disturbing the data to frustrate matching to external files.

In the intervening decade, traditional disclosure limitation techniques have come under significant academic scrutiny[1], resulting in improvements to traditional disclosure limitation techniques and in the development of new techniques based on mathematically formal definitions of privacy such as differential privacy[2]. As for microdata, a number of high-profile cases, some of which involve statistical agencies outside the United States, have demonstrated that it can be exceedingly difficult to evaluate the reidentification risk of a proposed microdata release.

With this background, the Confidentiality and Data Access Committee (CDAC) of the Federal Committee on Statistical Methodology (FCSM) hosted a one-day workshop on September 27, 2017, to present recent advances in the field of disclosure limitation. Over 125 people registered for the workshop, with 97 from the US Government, 12 from universities, and the remainder from industry. Of the government registrants, 30 were from the US Census Bureau, 13 from the Bureau of Labor Statistics, 9 from the Centers for Disease Control, and the remainder from other agencies. Approximately 80 people attended.

[1] For example, see "Revealing Information while Preserving Privacy," Dinur, I., and Nissim, K., PDOS 2003, June 9-12, 2003. San Diego, CA.

[2] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis, In Proceedings of the 3rd Theory of Cryptography Conference, 265–284.

Agenda

9:00 Welcome—Michael Hawes, Chair, CDAC

9:10 – 10:00 Disclosure Limitation Policy: Governance

9:10 IRBs and the Federal Wide Assurance Program
Jaime O. Hernandez, U.S. Dept. of Health and Human Services

9:40 What would a Federal Wide DRB Assurance Look like? A moderated Q&A
Jaime O. Hernandez, U.S. Dept. of Health and Human Services
Michael Hawes, U.S. Department of Education, Moderator

10:00 – 10:15 Break

10:15 – 11:45 Disclosure Limitation Policy: Differential Privacy

10:15 Differential Privacy: A Primer for a Non-Technical Audience
Kobbi Nissim, Georgetown University, Alexandra Wood, Harvard University

10:45 Differential Privacy and the 2020 Census
Simson L. Garfinkel, U.S. Census Bureau

11:15 Differential Privacy and the Federal Statistical Community? A moderated Q&A
Kobbi Nissim, Georgetown University
Alexandra Wood, Harvard University
Simson L. Garfinkel, U.S. Census Bureau
Darius Singpurwalla, National Science Foundation, Moderator

11:45 – 12:45 Lunch (on your own)

1:00 – 3:00 Advances in Disclosure Limitation Techniques 1

1:00 Can a Synthetic Data Approach Applied to High Risk Data Result in Usable Data with a Very Low Risk? Application to the Federal Employee Viewpoint Survey.
Taylor Lewis, U.S. Office of Personnel Management
Tom Krenzke, Westat

1:30 Data Hierarchies in Support of Disclosure Limitation
Shawn Merrill, Purdue University,
Keith Merrill, Brandeis University

2:00 Imputation as a Practical Alternative to Data Swapping
Saki Kinney, RTI International

2:30 Measuring Identification Risk in Microdata Release and its Control by Post-randomization
Cheng Zhang, George Washington University

3:00 – 3:15 Break

3:15 – 4:45 Advances in Disclosure Limitation Techniques 2

3:15 Synthetic Data for the American Community Survey
Rolando Rodríguez, U.S. Census Bureau
Michael H. Freiman, U.S. Census Bureau
Jerome P. Reiter, U.S. Census Bureau
Amy D. Lauger, U.S. Census Bureau

3:45 Evaluating the Consumer Expenditure Data Top-Coding Effects on Economic Models
Daniel Yang, U.S. Bureau of Labor Statistics
Daniell Toth, U.S. Bureau of Labor Statistics

4:15 Multivariate spatiotemporal modeling with applications to stroke mortality and data privacy
Harrison Quick, Drexel University

4:45 – 5:30 Town Hall and Closing Discussion

4:45 Where should we go from here?
Michael Hawes, U.S. Department of Education, Moderator,
Simson L. Garfinkel, U.S. Census Bureau,
Darius Singpurwalla, National Science Foundation

Slides from the sessions are available for download at
https://fcsm.sites.usa.gov/committees/cdac/2017-workshop/

Overview

The meeting began with a welcome from Michael Hawes, Chair, CDAC, who discussed the growing need for the federal statistical community to examine its disclosure limitation practices.

Session 1: Disclosure Avoidance Policy: Governance

Today many statistical agencies use *disclosure review boards (DRBs)* to oversee the release of statistical tables and microdata to assure that the requirements of various federal laws and regulations mandating the protection of confidential information be upheld. These laws include the Confidential Information Protection and Statistical Efficient Act (CIPSEA), The Family Educational Rights and Privacy Act (FERPA), the Health Insurance Portability and Accountability Act (HIPAA), and Titles 5, 13 and 26 of the US Code. These laws all exist to make sure that individuals and establishments are not identifiable in statistical publications, but they do not specify governance mechanisms for assuring that the legal requirements are maintained.

IRBs are similar to DRBs, except their purpose is to protect individuals before research is conducted, rather than before the results are published. IRBs are governed by the Common Rule (45 CFR 46), which draws its authority from the National Research Act. (PL 93-348). Unlike the statistical privacy laws, the Common Rule specifies in great detail the governance mechanisms that organizations using federal funds for research on humans must follow. In particular, the Common Rule specifies a system of certification and assurances that establishes the trustworthiness of an IRB to oversee human subject research, and both the record keeping and auditing requirements for maintaining an assurance.

With the idea that the IRB structure might be a useful model for the continued evolution of the DRB system, the opening session of the workshop featured a 30-minute tutorial by Jaime Hernandez of the US Dept. of Health and Human Services (HHS) about Institutional Review Boards (IRBs) and the approaches that HHS uses to ensure compliance with the Research Act and the Common Rule.

Following the tutorial, the attendees engaged in a 30-minute discussion regarding the current state of regulation of data releases. One topic that garnered significant interest was the question of using data sharing agreements for controlling the use of sensitive data by downstream data users. Another topic of concerns is that different federal agencies are using different statistical limitation standards and different data governance policies when they make their decisions. Some of these differences are explicitly discussed within the context of bilateral interagency agreements, while others are not.

Session 2: Disclosure Limitation Policy: Differential Privacy

Differential privacy is a mathematical definition for the privacy loss that individuals experience with their private information is used to create a data product. The term was coined by Cynthia Dwork in 2006. One of the key attributes of differential privacy is that it establishes a tradeoff between the accuracy of a data publication and the privacy loss to the individuals, and this tradeoff can be adjusted by the data publisher. Since 2006, hundreds of academic publications have developed a broad range of mechanisms that perform different kinds of data analysis and release in a manner that is differentially private, and both Apple and Google have incorporated differentially private data collection mechanisms into the iPhone and Chrome Browser, respectively.

The second session started with, Kobbi Nissim, one of the four inventors of differential privacy and now a professor at Georgetown University, and Alexandra Wood, a fellow at the Berkman Klein Center for Internet & Society at Harvard University, presenting "Differential Privacy: A Primer for a Non-Technical Audience."

Differential privacy is not a specific technique or algorithm, Nissim and Wood explained. Instead, differential privacy is a definition of privacy that "articulates a specific desiderata of an analysis: any information-related risk to a person should not change significantly as a result of that person's information being included, or not, in the analysis." That is, the output of an analysis should be similar depending on whether any specific person is included or not included in the data that were analyzed. The degree of similarity is gauged by the parameter ε (epsilon). The larger ε, the more privacy can be lost, and the more accurate the output. The accumulation of privacy loss can be bounded via the use of composition theorems for differential privacy, and hence ε can also be thought of as the *privacy loss budget*. Nissim and Wood said that in practice differential privacy is achieved by adding noise to data, either upon collection, or when the results are output. The particular choice of noise addition technique determines the accuracy of differentially private computations, and is task dependent. To conclude the presentation Nissim and Wood reviewed some of the challenges in making bringing differential privacy to use in real-life applications. One of these challenges is in uncovering how the use of differential privacy compares with the requirements of legal standards of privacy.

Following Nissim and Wood, Simson Garfinkel from the U.S. Census Bureau explained how the differential privacy is being incorporated into the 2020 Census. This is being done with a complex algorithm that models the responses of the entire nation and then assigns those response to different geographical units, from states, to counties, to census tracts, to block groups and finally to blocks. This model will then be used to create a set of privacy edits that are applied to the data collected from respondents. The resulting data will then be provided to the Bureau's tabulation system for the creation of publication tables. Unlike other uses of differential privacy, some of the values that the Bureau will be reporting are *invariant* and not subject to privacy edits, including the total count for each state (which is used for reapportioning the U.S. House of Representatives), and the total count for each block (which is used by states for restricting.)

Following the presentations there was a moderated Question and Answer session with the attendees asking questions of Nissim, Wood and Garfinkel, and having discussions among themselves regarding the use of differential privacy in modern statistical agencies.

Session 3: Advances in Disclosure Limitation Techniques 1

The first afternoon session looked at four approaches for improving privacy guarantees.

Partially Synthetic Data at OPM

Taylor Lewis from the US Office of Personnel Management discussed a research project in which they explored whether a partially synthetic dataset could be developed for external researchers to analyze data from OPM's Federal Employee Viewpoint Survey (FEVS). The FEVS is an annual survey of federal employees that "measures employees' perceptions of whether, and to what extent, conditions characterizing successful organizations are present in their agencies."[3] The primary consumers of the survey results are human resources managers who wish to assess how their workforce views compare with those of other agencies and the federal workforce as a whole. In 2016, the sample size was approximately 900,000, with over 80 agencies participating, and a response rate of just under 50%.

Respondent confidentiality is essential, otherwise some respondents might not answer the questions truthfully for fear of reprisal. The traditional disclosure approaches used in the survey have been top-coding, rounding, dropping especially sensitive variables, and creating separate files with separate sets of variables. Of particular concern are survey questions that identify a respondent's sexual orientation and race/ethnicity.

For this research, Lewis evaluated an approach for making a partially synthetic data set, in which some of the variables of each record would be from actual responses, while other variables would be synthesized. Synthesis was performed using the "synthpop" R

[3] https://www.fedview.opm.gov/

package and the "ctree" method using classification and regression trees (CART). Within a cell, values were synthesized randomly in proportion to their occurrence in the observed data.

Tom Krenzke from Westat discussed how a risk assessment was performed on the synthetic data. He discussed the estimated re-identification risk before and after synthetic approach was applied. The researchers concluded that the synthetic dataset dramatically reduced the disclosure risk, allowed more work units to be identified, allowed more demographic information to be included in the file, and it eliminated the need to have a separate "LGBT file." While results from the partially synthetic data did not match results generated with actual data exactly, most of the comparisons did not yield substantive differences.

The risk of re-identification was reduced from between 3% to 69% with a medium of 26% using conventional methods, to 0.43% overall, with 20 work units having a risk between 1.0% and 2.2%.

The summary findings presented were:

- Partially synthetic 2016 FEVS data does not produce perfect replications of the actual data, but results are reasonably close and devoid of any systematic biases
- Differences tend to zero as sample sizes increase
- Open question: is the extra noise a fair price to pay in exchange for more detailed demographic and work unit information, and dramatically reduced disclosure risk?

According to the speakers, while pleased with the reduced risk, OPM's management was not ready to adopt the partially synthetic approach until a more comprehensive assessment of the potential accuracy loss could be made.

Data Hierarchies in Support of Disclosure Limitation

In this presentation of preliminary work, Keith Merrill of Brandeis University and Shawn Merrill of Purdue University presented an approach that they are developing that leverages data hierarchies to allow the release of information using differential privacy with higher levels of accuracy than might be possible using the naïve application of traditional differential privacy mechanism.

In order to avoid leaking private data, differentially private mechanisms must be implemented in a manner that is data independent. This can pose a problem for many kinds of traditional data releases, where, for example, categories and hierarchies are usually decided by experts after looking at the data. For example, some age statistics reported by the US Census Bureau group population into age break of age 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 74, and 85 and over. On the other hand, other statistics are just reported for ages "65 and older." Such reporting, based on the data, can be problematic when using differential privacy

According to Merrill and Merrill, two important questions that a statistical agency must face when transitioning to the publication of high-dimensionality tables using differential privacy are:

1. Since every query now cuts into our privacy budget, when should we create a new table by combining smaller ones, as opposed to querying the data again?
2. How do we automate the process of hierarchy generation, to remove/lessen the need for SMEs to be savvy in the privacy literature?

With respect to the first question, combining results does not count against the budget, but does have much larger variance in the answers than running another query.

Imputation as a Practical Alternative to Data Swapping

In this presentation, Saki Kinney from RTI International discussed the use of synthetic data ("imputation") as an alternative to swapping as a confidentiality protection mechanism.

According to Kinney, data swapping is used by several agencies for "demographic, lower risk datasets." Swapping is attractive because it precisely preserves marginal distributions, but it distorts relationships between swapped and unswapped values. There are also few publicly available swapping implementations, and organizations that use swapping cannot reveal the swapping rates that they use, because revelations about the swapping procedures can compromise the privacy protection that swapping may provide. Researchers have also found that even very low swapping rates can negatively impact data quality.

One approach for using "synthetic data in a swapping world," Kinney said, is to identify records and variables that would normally be swapped, and "instead of swapping, replace values with (single) imputations." The advantage of this approach, Kinney said, is that imputation provides a model-based, flexible, intuitive alternative to swapping.

To demonstrate this approach, Kinney performed an experiment with the National Science Foundation's National Survey of College Graduates Public Use File. Treating this dataset as a confidential dataset, the procedure imputed 18.9% of records for 7 variables. Ten graphs were presented based on both the original and the partially synthetic datasets, and visual inspection showed the two datasets to be quite similar.

A problem with this approach, Kinney readily admitted, is that the results will not be formally private as long as the disclosure protection relies on the fact that it is uncertain which records have been perturbed and which have not. But this approach has the advantage that it provides greater transparency and flexibility compared to swapping, and can be used today.

Measuring Identification Risk in Microdata Release and its Control by Post-randomization

Respondent privacy can be protected by adding noise to released data. In the fourth presentation of this session, Cheng Zhang proposed a new approach for measuring identity disclosure called *identification risk* (IR). With this, he proposed a statistical perturbation method called *Inverse Frequency Post-Randomization* (IFPR) that allows control over identification risk while minimizing data quality loss. The method requires that the data publisher identify *key variables* and then considers the identification risk as resulting from matches on these key variables. Other variables are considered *non-key variables* and do not pose an identification risk in this model.

Session 4: Advances in Disclosure Limitation Techniques 2

The second afternoon session consisted of three more research presentations.

Synthetic Data for the American Community Survey

Rolando Rodríguez presented work being done at the U.S. Census Bureau to create a synthetic dataset that models the American Community Survey (ACS), with the ultimate goal of creating an ACS data release that is formally private.

The ACS is a survey of approximately 2.3 million housing units a year. More than a 1000 tables are distributed for every Census block group. Roughly two-thirds of samples are released as a public use microdata, but with coarsened geography.

Rodríguez explained that the methods being developed to make the decennial census formally private will not work for the ACS, because the ACS has more characteristics for housing units and people, and the ACS has complex survey weights. So instead, the plan is to create a synthetic data set, and then to make the synthesis formally private by adding noise to the model. Currently the team is using two models for the synthesis: classification and regression trees (CART) to synthesize factors and counts, and then a linear regression to synthesize (rounded) continuous variables. A danger in using CART, however, is that trees with too many leaves can lead to overfitting, which could directly reproduce respondent data and lead to privacy violations. And even if the project works, a risk is that users might perform an analysis on the ACS public-use microdata that was not considered during the synthesis.

Key questions that remain on this project include:

- Can we use formal privacy methods on some subset of the variables?
- Can we make current methods formally private?
- How do we account for survey weights?
- How do results look after placing housing units in sub-state geographies?
- How can we leverage alternate data sources (administrative records)?

Evaluating the Consumer Expenditure Data Top-Coding Effects on Economic Models

Top-coding is a common disclosure limitation technique used to hide outliers. A problem with top-coding is that it can cause skewed results by eliminating high values that might otherwise influence means and variances. This paper evaluated the impact of top-coding microdata for the Bureau of Labor Statistics Consumer Expenditure study.

One way to minimize such impact on accuracy is to use a top-code value that is higher than the cut-off value. For example, if there are 19 people in a geographical area whose salary is reported, and the majority are under $30K, but there are also people with salary of $50K, $75K, $125K, $275K and $540K, it may be necessary to top-code those with a salary over $75K. The salaries could be top-coded at the cut-off value of $75K, or they could be top-coded at the average of the salaries that are cut-off, which would be $253K.

To determine the impact of top-coding, Yang and Toth examined Income Elasticity of Demand using public release Consumer Expenditure Data from 2008, considering expenditure outcomes for utilities, domestic services, transportation, shelter, medical supplies, major appliances, other vehicle, and new cars and trucks. They found no difference using a model with log linear regression between models created with confidential and top-coded data, but they did find a difference using the model with logistic regression. This translates into some differences in income elasticity of demand for some expenditures.

Even though certain expenditures are infrequent but they still are of interest to economists and industry. Therefore, the BLS the program office may be able to come up with a warning to economists or researchers on the differences of economics measurements due to top-coding and acceptable threshold.

Multivariate spatiotemporal modeling with applications to stroke mortality and data privacy

In joint work with Lance Waller (Emory) and Michele Casper (Centers for Disease Control), Harrison Quick presented a proposal for creating synthetic data for an important but privacy sensitive CDC data release.

The CDC operates a nationwide epidemiological surveillance network, and disseminates this information in articles and reports. It also releases epidemiological data via CDC

WONDER (https://wonder.cdc.gov), an on-line public health information system. Currently, CDC WONDER suppresses many low-incidence observations. This is problematic, as it results in demographic aspects of many diseases being understudied.

The presentation used the example of stroke mortality. Using current National Center for Health Statistics (NCHS) guidelines, instances with counts less than 10 must be suppressed, which leads to suppressing 70% of data when performing an analysis of stroke by county, age, and year of stroke. Using a synthetic data model, the researchers were able to impute cells with counts below 10. Once this is done, the researchers can measure the disclosure risk of the resulting data, as well as the utility, which they define as the degree to which the resulting data fits a Poisson model. As a result of using the imputation model, the majority of the cells that had been suppressed can instead be provided, but with synthetic data.

The goal of the project is to be able to generate a series of synthetic one-offs for individual analysis, rather than to generate a “Synthetic CDC WONDER.”

A key limitation of this approach, the Quick stated, is that it is unclear how to connect this approach with differential privacy, and it is not practical situations where there are many subgroups unless many assumptions are made. Another question is whether relationships that are not included in the model will be preserved in the synthetic data.

Where do we go from here?

Following the presentations, there was a brief open discussion about the next steps for statistical agencies in improving the state-of-the-art of disclosure limitation techniques. Overall, those in attendance expressed that moving to formal privacy techniques requires sophistication and skills that are not readily available. There were also concerns that data users might not be sufficiently skilled to make use of products that incorporate formal privacy. Many agencies are waiting to see the experience of the U.S. Census Bureau in incorporating formal privacy in the 2020 census.