RESEARCH REPORT SERIES
(Statistics #2009-09)

Recent Work on the Microdata Analysis System at the Census Bureau

Jason Lucero
Lisa Singh[1]
Laura Zayatz

[1] Also Georgetown University, Computer Science Department, Washington, DC 20057

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: November 16, 2009

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau

Recent Work on the Microdata Analysis System at the Census Bureau

Jason Lucero, Lisa Singh[2], and Laura Zayatz
U.S. Census Bureau[1], Commerce/Census/SRD/5K114F, 4600 Silver Hill Road, Washington, DC 20233-9100,
Jason.Lucero@census.gov

Abstract
The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code which promises confidentiality to its respondents. The agency also has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality information as possible without violating the pledge of confidentiality. This paper discusses a Microdata Analysis System (MAS) that is under development at the Census Bureau. The system is designed to allow data users to perform various statistical analyses (for example, regressions, table generations, generation of correlation coefficients) of survey and census data without seeing or downloading the actual underlying confidential microdata. This paper begins with an overview of the Microdata Analysis System and discusses the confidentiality rules currently implemented in the system. The remainder of the paper will focus on a recently evaluated *Drop q Rule* and a cutpoint generation program. These are used to protect the confidentiality of results generated from the system while still maintaining data quality and utility. We will then conclude with a brief discussion about future work on the MAS.

Key Words: Disclosure Avoidance, Confidentiality, Data Dissemination, Remote Access, Differencing Attack, Sub-sampling

1. Introduction

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code. This prevents the Census Bureau from releasing any data "...whereby the data furnished by any particular establishment or individual under this title can be identified." In addition to Title 13, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) requires the protection of information collected or acquired for exclusively statistical purposes under a pledge of confidentiality. In addition, the agency has the responsibility of releasing data for the purpose of statistical analysis. In common with most national statistical institutes, our goal is to release as much high quality data as possible without violating the pledge of confidentiality (Duncan, Keller-McNulty, and Stokes, 2003; Kaufman, Seastrom, and Roey, 2005). We apply disclosure avoidance techniques prior to publicly releasing our data products to protect the confidentiality of our respondents and their data (Willenborg and de Waal, 2001). This paper discusses a Microdata Analysis System (MAS) that is under development at the Census Bureau. The system is designed to allow data users to perform various statistical analyses (for example, regressions, table generations, generation of correlation coefficients) of survey and census data without seeing or downloading the actual underlying confidential microdata. We begin by answering some frequently asked questions about the MAS. We then

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

[2] Georgetown University, Computer Science Department, Washington, DC 20057

discuss the current state of the system, including which data sets and types of statistical analyses are included, and the confidentiality rules used to protect data products generated from the MAS. We then discuss some of the recent work that has been done on the MAS. In particular, we focus on a *Drop q Rule* and a cutpoint generation program. We end with remarks on future work.

2. Answers to Frequently Asked Questions about the Microdata Analysis System

2.1 Why Do We Need a MAS?

There are several reasons to develop a Microdata Analysis System. One reason is to allow data users to perform statistical analyses on the actual confidential microdata instead of our public use microdata files, which have been modified to protect the confidentiality of our respondents. The Census Bureau conducts reidentification studies on its public use microdata files. In these studies, we attempt to link outside files that have identifiers on them to our public use files. We have found and fixed a few small problems, but there is a concern that more problems will arise in the future because more and more data are becoming publicly available on the internet, and more people are using record linkage software and data mining in an effort to increase the amount of information they can work with. As a result, data users are worried that we may have to cut back on the detail in our public use files and use more data perturbation techniques to protect them.

A second (related) reason for developing the MAS is not to allow data users to access new information, but to allow data users to access more detailed, accurate information than what is currently available on our public use files. For example, perhaps the data that can be accessed through the MAS could identify smaller geographic areas and show more detail in variable categories and tail ends of distributions that are normally not shown on public use files. One of our goals for the MAS is to allow access to as much high quality data as possible (Weinberg et al., 2007; Rowland and Zayatz, 2001). In addition, the MAS would allow users to access to data that is not typically released in public use microdata files. This would include data on establishments and linked data sets that include both demographic and establishment data.

2.2 What Data Sets and What Types of Analyses Will be Available on the MAS?

The ultimate goal is to include any and all data sets and any and all types of analyses. We will begin with data from demographic surveys and the decennial censuses. We would like to add establishment survey and census data as well as linked data sets. We will begin with regressions, cross-tabulations, and correlation coefficients, and add other applications in the future. See section 3 for what is included in the prototype.

2.3 Who Might Want to Use the MAS and Will It Cost Anything?

The MAS will be used by people with needs for fairly simple statistical analyses (news media, some policy makers, teachers and students). We understand that some users feel the need to use the underlying microdata for more exploratory data analysis. However, due to confidentiality concerns, users will have to continue to use the public use files (although they may not offer the detail that one might get through the MAS) or the Research Data Centers (although this is not as inexpensive or easy to use as the MAS) for their exploratory data analysis needs.

A final decision on cost has not yet been made; however, the current plan is to offer this as a free service through the Census Bureau's DataFERRETT (Chaudhry, 2007).

2.4 Will the Census Bureau Keep Track of Who Uses the MAS and What Queries Are Submitted?

We are currently investigating the legality of doing this. There are at least two reasons why we would want to. First, we want to see how people are using the system, so that we can make modifications and enhancements to improve the user experience.

The second reason would be for disclosure avoidance purposes. These data may be useful to help identify disclosure risks arising from multiple queries to the system. The system is meant to do all disclosure avoidance by itself through confidentiality rules and restrictions on the underlying datasets. There will be no humans monitoring the system. Instead, we will write software that prevents users and automated robot programs from bombarding the system with large numbers of queries. We will not try to determine if different users are colluding with multiple queries. There will be no modification of data on the fly.

3. An Alpha Prototype of the Microdata Analysis System

The Census Bureau contracted with Synectics to develop an alpha prototype of the MAS. It was written in SAS. We also contracted with Jerry Reiter of Duke University to help us develop the confidentiality rules and Steve Roehrig of Carnegie Mellon University to help us test the confidentiality rules. Some rules were modified as a result of the testing.

3.1 Data Sets and Types of Statistical Analyses

The data sets included in the prototype were the Current Population Survey (CPS) March 2000 Demographic Supplement and the 2005 American Community Survey (ACS). The types of statistical analyses available in the prototype were cross-tabulations, generation of correlation coefficients, ordinary least squares (OLS) regression, binary logistic regression and multinomial logistic regression.

3.2 Current Confidentiality Rules For Universe Formation Within the Alpha MAS Prototype.

The confidentiality rules discussed in this section and in section 4 are quite complex. This paper gives a brief overview of them. More detail can be found in Lucero[2] (2009).

The MAS software is programmed with several confidentiality rules and procedures that uphold disclosure avoidance standards. The purpose of these rules and procedures is to prevent data intruders from exploiting the system by submitting multiple statistical queries for the purpose of recreating individual confidential microdata records. Currently, the alpha prototype implements rules on the formation of universes (subpopulations) and rules for OLS regression, binary logistic regression, and multinomial logistic regression. The alpha prototype does not implement any confidentiality rules for cross-tabulations or for the generation of correlation coefficients. Work is currently underway at the Census Bureau to develop a beta prototype, that will include confidentiality rules for cross-tabulations and correlation coefficients, as well as confidentiality rules that required modification due to previous testings. This recent work will be discussed further in sections 4 and 5.

All variables used for universe formations are categorical recodes of the actual raw variables found in the underlying microdata. Raw categorical variables, as well as their corresponding category level bins, are coded directly into the metadata. Raw numerical variables are presented to the user as recoded categorical variables based on output from a cutpoint program. This cutpoint program bins numerical values to further confidentiality protection. Details of the cutpoint program will be discussed in section 5. These recoded variables are used for universe formations only. All statistical analyses performed on the MAS use the raw variable values from the underlying microdata.

Universe formation on the MAS is performed through an implicit table server. To form a universe, users would first select *m* recoded variables, then select up to *j* observed bin levels for each of the *m* recoded variables. Currently, users can define a universe using no more than $m = 4$ variables, and may select no more than $j = 8$ observed bin levels for each variable. The MAS would then generate an *m*-way table of counts for the *m* recoded variables used to define the universe. A universe query on the MAS can be thought of as a request for a set of cell counts from the *m*-way table of counts.

For example, suppose the user would like to perform a statistical analysis on the following universe:

$P_1(99) = [\textit{gender} = \text{female and } \$28{,}501 \leq \textit{income} \leq \$39{,}500]$
OR
$P_2(49+11) = [\textit{gender} = \text{male and } \$62{,}001 \leq \textit{income} \leq \$120{,}000]$

This universe is derived from the set of yellow and blue cells from Table 3.2.1, a two-way table of counts for *gender* by *income*. The yellow cell represents the first piece of the full universe, $P_1(99)$. The blue cells represent the second piece of the full universe, $P_2(49+11)$. Note that there are 99 total observations in P_1 and 60 total observations in P_2. Furthermore, since no cell counts are shared among P_1 and P_2, we would call this a *disjoint universe*, U^d. There is a total of 159 observations in the full disjoint universe: $U^d(159) = P_1(99)$ or $P_2(49+11)$.

	income							
gender	$0 to $28,500	$28,501 to $39,500	$39,501 to $45,000	$45,001 to $53,500	$53,501 to $62,000	$62,001 to $70,500	$70,501 to $120,000	Total
male	20	97	49	92	38	49	11	356
female	26	99	42	64	45	37	8	321
Total	46	196	91	156	83	86	19	677

Table 3.2.1-Example of a disjoint universe.

As for another example, suppose the user requests the following universe on the MAS:

$P_1(321)$ = [*gender* = female]
OR
$P_2(86+19)$ = [\$62,001 $\leq$ *income* $\leq$ \$120,000]

This universe is also derived from the same two-way table of counts for *gender* by *income*, as shown in Table 3.2.2. The yellow and green cells represents the first piece of the full universe, $P_1(321)$. The blue and green cells represent the second piece of the full universe, $P_2(86+19)$. The green cells represent the shared cell counts among P_1 and P_2. That is, the intersection of P_1 and P_2 is non-empty, and contains 37+8 = 45 total observations: $P_1(321) \cap P_2(86+19) = I(37+8)$. Since cell counts are shared among P_1 and P_2, we would call this a *joint universe*. There are 321+86+19 – 37 – 8 = 381 total observations within the full joint universe: $U^j(381) = P_1(321)$ or $P_2(86+19)$

	income							
gender	\$0 to \$28,500	\$28,501 to \$39,500	\$39,501 to \$45,000	\$45,001 to \$53,500	\$53,501 to \$62,000	\$62,001 to \$70,500	\$70,501 to \$120,000	Total
male	20	97	49	92	38	49	11	356
female	26	99	42	64	45	37	8	321
Total	46	196	91	156	83	86	19	677

Table 3.2.2-Example of a joint universe.

Before users are allowed to perform any statistical analyses on their chosen universe, their universe must pass the following two rules:

- *No Marginal 1 or 2 Rule:* The *m*-way table, from which the universe is derived, cannot contain any (*m*-1) dimensional marginal totals equal to 1 or 2.

- *75 Rule*: The universe must contain at least 75 observations.

If at least one of these two rules fail, then the MAS rejects the user's universe query, and prompts the user to modify his universe selection. For any universe derived from an *m*-way table of counts, the *No Marginal 1 or 2 Rule* ensures that all (*m*-1) dimensional marginal totals are not equal to 1 or 2. If at least one (*m*-1) dimensional marginal total equals 1 or 2, then no universes can be derived from that particular *m*-way table. The application of the *75 Rule* is dependent if the universe is disjoint or joint. If a universe is disjoint, then no cell counts are shared among any of its pieces, and each piece is checked separately for the *75 Rule*. If a universe is joint, then there are at least some cell counts shared among two or more of its pieces, and each of these shared cell counts must be checked for the *75 Rule*. Furthermore, all cutpoint bins are combined within each piece or within each set of shared cell counts in order to test the *75 Rule*.

For example, to apply the *75 Rule* for the disjoint universe $U^d(159) = P_1(99)$ or $P_2(49+11)$, derived from Table 3.2.1, the *75 Rule* must be tested separately for $P_1(99)$ and $P_2(49+11)$. Both P_1 and P_2 must each contain 75 or more observations. The *75 Rule* is satisfied for P_1 since $99 \geq 75$. Since *income* is a categorical recode, the cutpoint bins of *income* are combined to test the *75*

Rule for P_2. However, since 49+11 = 60 < 75, the *75 Rule* fails for P_2 and the MAS would reject the user's universe query.

To apply the *75 Rule* for the joint universe $U^j(384) = P_1(321)$ or $P_2(86+19)$, derived from Table 3.2.2, the *75 Rule* must be tested separately for the shared cell counts contained in the non-empty intersection $I(37+8) = P_1(321) \cap P_2(86+19)$. The total number of observations within this non-empty intersection must contain 75 or more observations. Once again, the income cutpoint bins must be combined to test the *75 Rule* within the non-empty intersection I(37+8). Since 37+8 = 45 < 75, the *75 Rule* fails for I(37+8), and the MAS would reject the user's universe query. Further details and examples of testing the *75 Rule* for disjoint and joint universe types can be found in Lucero[2], (2009).

3.3 Current Confidentiality Rules For Regression Analyses within the Alpha MAS Prototype.

In addition, the MAS implements other confidentiality rules for regression analyses. Again, we will provide only a brief overview of these rules. More detail can be found in Lucero[3] (2009). For example, no more than 20 independent variables may be selected for any regression model. Furthermore, since there are some transformations that could deliberately emphasize outliers (Gomatam et al., 2004), the set of transformations available to the user is limited to a predetermined list.

As shown in Reznek (2003) and Reznek and Riggs (2004), any fully interacted regression model that contains only dummy variables as predictor variables can pose a disclosure risk. Therefore, users are restricted to include only two-way and three-way interaction terms within their regression model. In addition, each predictor dummy variable must pass a minimum size threshold, or it will be absorbed into the intercept term along with the dummy variable that represents the reference category level.

Depending on the type of regression analysis, the MAS will check the values for some summary statistics prior to passing back the estimated regression coefficients back to the user. For example, for OLS regression, the MAS checks and ensures that R^2 is not too close to 1. If R^2 is too close to 1, then the fitted regression model fits the data too well, and users could use the values of the estimated coefficients to obtain accurate predictions of the response variable given known values of the predictor variables (Reiter, 2004). If R^2 is too close to 1, then the MAS will not output any regression results back to the user. If R^2 is not too close to 1, then the MAS will output the estimated regression coefficients, as well as the ANOVA table, to the user without restriction.

The MAS never outputs the actual residual values back to the user, since the real residual values can be easily manipulated to determine the actual values of the dependent variable. All diagnostic residual plots on the MAS are based on synthetic residuals vs. synthetic fitted values, which are designed to mimic the patterns shown in the actual scatter-plot of the real residuals vs. real fitted values. These synthetic residual scatter-plots allow the user to check the fit of their submitted regression model, without the disclosure of the actual residual values (Reiter, 2003).

4. Recent Work on the Microdata Analysis System: The *Drop q Rule*

In this section, we present a brief overview of the recent evaluation of the Microdata Analysis System's universe subsampling routine known as the *Drop q Rule*. A more in depth evaluation can be found in Lucero[1] (2009).

4.1 Differencing Attack Disclosures

While the *75 Rule* ensures that the universe data set, U(*n*), meets a minimum size requirement, it does not prevent differencing attack disclosures. A *differencing attack* combines the statistical results obtained from two similar universe queries in an attempt to disclose an individual's confidential microdata record. To perform a differencing attack disclosure, a data intruder would first create two similar universes on the MAS:

U(*n*): A universe with n total observations.

U(*n*-1): A universe with the exact same n observations as U(*n*), less one observation.

The difference U(*n*) – U(*n*-1) = U(1), where U(1) is a universe that contains only one unique observation. Suppose a data intruder then requests two similar two-way tables of counts for *gender* by *race* on the MAS: T[U(*n*)] from U(*n*), and T[U(*n*-1)] from U(*n*-1), as shown in Figure 4.1.1.

T[U(*n*)]	*race*				
gender	White	Black	Asian	Other	Total
Male	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
Female	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	

-

T[U(*n*-1)]	*race*				
gender	White	Black	Asian	Other	Total
Male	n_{11} - 1	n_{12}	n_{13}	n_{14}	$n_{1.}$ - 1
Female	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
Total	$n_{.1}$ -1	$n_{.2}$	$n_{.3}$	$n_{.4}$	

=

T[U(1)]	*race*				
gender	White	Black	Asian	Other	Total
Male	1	0	0	0	1
Female	0	0	0	0	0
Total	1	0	0	0	

Figure 4.1.1-Example of a differencing attack performed on two similar two-way tables.

Since U(*n*) and U(*n*-1) only differ by one unique observation, T[U(*n*-1)] will be exactly the same as T[U(*n*)], less one cell count. This sensitive cell is shaded in Figure 4.1.1. The data intruder then performs the following differencing attack on these similar two-way tables: T[U(*n*)] – T[U(*n*-1)] = T[U(1)]. The resulting table, T[U(1)], is a two-way table of counts of *gender* by *race*, which contains a cell count of 1 in the cell that represents white males. By performing a differencing attack T[U(*n*)] – T[U(*n*-1)] = T[U(1)], the data intruder has disclosed that the one unique observation contained in U(1) = U(*n*) – U(*n*-1) is a white male.

4.2 The *Drop q Rule*

To help guard against differencing attack disclosures, the MAS implements a subsampling routine called the *Drop q Rule*. After the universe U(*n*) passes the *No Marginal 1 or 2 Rule* and the *75 Rule*, *q* observations are randomly removed from the U(*n*) data set to yield a new subsampled universe data set, U(*n*-*q*), where *q* << *n*. If the same U(*n*) is selected again by the same user, or by a different user, then the exact same *q* observations are dropped from U(*n*) to yield the same subsampled U(*n*-*q*) as before. Currently, the value of *q* is fixed for all universe queries on the MAS and does not change with respect to the universe size *n*.

On the MAS, every statistical analysis is performed on the subsampled U(*n*-*q*) data set, and not on the original U(*n*) data set. Therefore, if a data intruder attempts to perform a differencing attack on two similar two-way tables of *gender* by *race*, T[U(*n*)] – T[U(*n*-1)], as shown in Figure 4.2.1, he is actually performing a differencing attack of T[U(*n*-*q*)] – T[U(*n*-1-*q*)] as shown in Figure 4.2.2.

T[U(*n*)]	*race*			
gender	White	Black	Asian	Other
Male	n_{11}	n_{12}	n_{13}	n_{14}
Female	n_{21}	n_{22}	n_{23}	n_{24}

–

T[U(*n*-1)]	*race*			
gender	White	Black	Asian	Other
Male	$n_{11} – 1$	n_{12}	n_{13}	n_{14}
Female	n_{21}	n_{22}	n_{23}	n_{24}

Figure 4.2.1- The differencing attack T[U(*n*)] - T[(U(*n*-1)].

T[U(*n*-*q*)]	*race*			
gender	White	Black	Asian	Other
Male	$n_{11} – X_{11}$	$n_{12} – X_{12}$	$n_{13} – X_{13}$	$n_{14} – X_{14}$
Female	$n_{21} – X_{21}$	$n_{22} – X_{22}$	$n_{23} – X_{23}$	$n_{24} – X_{24}$

–

T[U(*n*-1-*q*)]	*race*			
gender	White	Black	Asian	Other
Male	$n_{11} – 1 – Y_{11}$	$n_{12} – Y_{12}$	$n_{13} – Y_{13}$	$n_{14} – Y_{14}$
Female	$n_{21} – Y_{21}$	$n_{22} – Y_{22}$	$n_{23} – Y_{23}$	$n_{24} – Y_{24}$

Figure 4.2.2- The differencing attack T[U(*n*-*q*)] - T[(U(*n*-1-*q*)].

In Figure 4.2.2, T[U(*n*-*q*)] is a two-way table for *gender* by *race*, based on the subsampled U(*n*-*q*) data set, and T[U(*n*-1-*q*)] is a two-way table for *gender* by *race* based on the independently subsampled U(*n*-1-*q*) data set. The random removal of *q* observations from U(*n*) to yield U(*n*-*q*) is equivalent to removing *q* cell counts at random from the original two-way table T[U(*n*)] to yield a new subsampled two-way table of *gender* by *race*, T[U(*n*-*q*)]. Similarly, the random removal of *q* observations from U(*n*-1) to yield U(*n*-1-*q*) is equivalent to removing *q* cell counts at random from the original two-way table T[U(*n*-1)] to yield a subsampled two-way table T[U(*n*-1-*q*)]. The random variables X_{ij} give the number of counts that were randomly removed from each cell in T[U(*n*-*q*)], while the random variables Y_{ij} give the number of counts that were randomly removed from each cell in T[U(*n*-1-*q*)], where $0 \leq X_{ij} \leq q$, $0 \leq Y_{ij} \leq q$, $\Sigma_i \Sigma_j X_{ij} = q$ and $\Sigma_i \Sigma_j Y_{ij} = q$. Since T[U(*n*-*q*)] and T[U(*n*-1-*q*)] are based on two independently subsampled universe data sets, the resulting table T[U(1)] = T[U(*n*-*q*)] – T[U(*n*-1-*q*)] may or may not yield a successful disclosure of *gender* and *race* for the one unique observation contained in U(1).

4.3 An Evaluation of the *Drop q Rule*

On the MAS alpha prototype, $q = 2$ regardless of the size of the original U(*n*) data set. We believed that higher values of *q* would yield lower probabilities of obtaining successful disclosures from the *m*-way table T[U(1)] = T[U(*n*-*q*)] – T[U(*n*-1-*q*)]. However, we observed that the distribution of cell proportions within the original *m*-way table T[U(*n*)], on which the data intruder attempts to perform a differencing attack, played a major role in determining the effectiveness of the *Drop q Rule*, not just the value of *q* itself.

For example, using the same similar two-way tables of *gender* by *race*, T[U(n)] and T[U(n-1)], as shown in Figure 4.2.1, when q observations are removed at random from U(n) to yield U(n-q), the observed number of counts x_{ij} that were randomly removed from each cell_{ij} in resulting two-way table T[U(n-q)] (as shown in Figure 4.2.2), is dependent on the proportion of counts π_{ij} within each cell_{ij} in the original two-way table T[U(n)]. If we think of these cell proportions as probabilities, then, for large values of n, the random variables $X_{11},\ldots,X_{24}$ in T[U(n-q)] follow an approximate multinomial distribution with parameters q, $\pi_{11},\ldots,\pi_{24}$:

$$P\left(X_{11}=x_{11},...,X_{24}=x_{24}\mid q,\pi_{11},...,\pi_{24}\right)=\frac{q!}{x_{11}!\cdots x_{24}!}\pi_{11}^{x_{11}}\cdots\pi_{24}^{x_{24}}$$

Similarly, when q observations are removed at random from U(n-1) to yield U(n-1-q), the observed number of counts y_{ij} that were randomly removed from each cell_{ij} in the resulting two-way table of *gender* by *race*, T[U(n-1-q)], is also dependent on the cell proportions within the original two-way table T[U(n-1)]. However, for large values of n, the cell proportions in T[U(n-1)] are approximately equal to the cell proportions π_{ij} in the original two-way table T[U(n)]. Therefore, the random variables $Y_{11},\ldots,Y_{24}$ T[U(n-1-q)] also follow an approximate multinomial distribution with same parameters q, $\pi_{11},\ldots,\pi_{24}$:

$$P\left(Y_{11}=y_{11},...,Y_{24}=y_{24}\mid q,\pi_{11},...,\pi_{24}\right)=\frac{q!}{y_{11}!\cdots y_{24}!}\pi_{11}^{y_{11}}\cdots\pi_{24}^{y_{24}}$$

Since U(n-q) and U(n-1-q) are subsampled independently, T[U(n-q)] and T[U(n-1-q)] are two independently subsampled tables, and the approximate joint probability of removing $X_{11} = x_{11}$, $\ldots,X_{24} = x_{24}$ cell counts at random from T[U(n-q)] and removing $Y_{11} = y_{11},\ldots,Y_{24} = y_{24}$ cell counts at random from T[U(n-1-q)] is:

$$\text{(4.3.1)}\quad P\left(X_{11}=x_{11},...,X_{24}=x_{24}\mid q,\pi_{11},...,\pi_{24}\right)P\left(Y_{11}=y_{11},...,Y_{24}=y_{24}\mid q,\pi_{11},...,\pi_{24}\right)=$$
$$\left(\frac{q!}{x_{11}!\cdots x_{24}!}\right)\left(\frac{q!}{y_{11}!\cdots y_{24}!}\right)\pi_{11}^{x_{11}+y_{11}}\cdots\pi_{24}^{x_{24}+y_{24}}\text{, where } \Sigma_i\Sigma_j X_{ij} = q,\ \Sigma_i\Sigma_j Y_{ij} = q,\ \Sigma_i\Sigma_j \pi_{ij} = 1.$$

The PMF in 4.3.1 gives us the approximate joint probability of obtaining a subsampled table T[U(n-q)] from T[U(n)] and a subsampled table T[U(n-1-q)] from T[U(n-1)], where $X_{ij} = x_{ij}$ counts were removed at random from each cell_{ij} in T[U(n-q)] and $Y_{ij} = y_{ij}$ counts were removed at random from each cell_{ij} in T[U(n-1-q)].

It was observed that a successful disclosure of *gender* and *race* from the differencing attack of T[U(1)] = T[U(n-q)] - T[U(n-1-q)] could occur only if $x_{ij} = y_{ij}$, for all i, j. That is, the x_{ij} counts that were removed at random from each cell_{ij} in T[U(n-q)] must exactly match the y_{ij} counts that were removed at random from each cell_{ij} in T[U(n-1-q)]. Therefore, the joint approximate probability of obtaining two such subsampled tables T[U(n-q)] and T[U(n-1-q)], where the exact same $x_{ij} = y_{ij}$ counts were removed at random from each cell_{ij} in both T[U(n-q)] and T[U(n-1-q)], is:

(4.3.2) $$P\left(X_{11}=x_{11},\ldots,X_{24}=x_{24}\mid q,\pi_{11},\ldots,\pi_{24}\right)=\left(\frac{q!}{x_{11}!\cdots x_{24}!}\right)^2\pi_{11}^{2x_{11}}\cdots\pi_{24}^{2x_{24}},$$

where $\Sigma_i\Sigma_j X_{ij} = q$, and $\Sigma_i\Sigma_j \pi_{ij} = 1$.

As a result, 4.3.2 gives us the approximate probability of obtaining a successful disclosure of *gender* and *race* from a single differencing attack T[U(1)] = T[U(*n*-*q*)] - T[U(*n*-1-*q*)]. If we sum 4.3.2 over all possible sequences of $x_{11},\ldots x_{24}$, such that $\Sigma_i\Sigma_j X_{ij} = q$, we get 4.3.3, the total approximate probability of obtaining a successful disclosure of *gender* and *race*, for all possible differencing attacks of T[U(1)] = T[U(*n*-*q*)] - T[U(*n*-1-*q*)]:

(4.3.3) $$\sum_{x_{11},\ldots,x_{24}}^{x_{11}+\ldots+x_{24}=q}\left(\frac{q!}{x_{11}!\cdots x_{24}!}\right)^2\pi_{11}^{2x_{11}}\cdots\pi_{24}^{2x_{24}}$$

In general, given any two similar universe data sets U(*n*) and U(*n*-1), and their independently subsampled universe data sets U(*n*-*q*) and U(*n*-1-*q*), if a data intruder requests the same *m*-way table of counts T[ ] for both U(*n*-*q*) and U(*n*-1-*q*), where T[ ] contains Λ total cells, and performs the differencing attack T[U(*n*-*q*)] – T[U(*n*-1-*q*)] = T[U(1)] as an attempt to disclose all *m* observed categorical variables for the one unique observation contained in U(1) = U(*n*) – U(*n*-1), then the total approximate probability of obtaining a successful disclosure from T[U(1)] for all *m* observed categorical variables is given by:

(4.3.4) $$\sum_{x_1,\ldots,x_\Lambda}^{x_1+\ldots+x_\Lambda=q}\left(\frac{q!}{x_1!\cdots x_\Lambda!}\right)^2\pi_1^{2x_1}\cdots\pi_\Lambda^{2x_\Lambda}$$

where $\pi_1,\ldots,\pi_\Lambda$ are the cell proportions contained in the *m*-way table T[U(*n*)], U(*n*) is the original data set, $\pi_1 + \cdots + \pi_\Lambda = 1$, and the summation in 4.3.4 is taken over all possible sequences of $x_1,\ldots,x_\Lambda$ such that $x_1 + \cdots + x_\Lambda = q$.

Using the function NMinimize in Mathematica, we set equation 4.3.4 as a function of $\pi_1,\ldots,\pi_\Lambda$, and then performed several non-linear optimization routines to minimize 4.3.4 subject to the constraint $\pi_1 + \cdots + \pi_\Lambda = 1$ for $\Lambda = 2,\ldots,9$, and different values of *q*. It was found that the minimum total approximate probability of obtaining a successful disclosure was achieved when $\pi_1 = \pi_2 = \cdots = \pi_\Lambda = 1/\Lambda$, regardless of the value of *q*. Setting $\pi_1 = \cdots = \pi_\Lambda = 1/\Lambda$ in equation 4.3.4, the minimum total approximate probability of obtaining a successful disclosure from T[U(*n*-*q*)] – T[U(*n*-1-*q*)] = T[U(1)] is:

(4.3.5) $$\frac{\displaystyle\sum_{x_1,\ldots,x_\Lambda}^{x_1+\ldots+x_\Lambda=q}\left(\frac{q!}{x_1!\cdots x_\Lambda!}\right)^2}{\Lambda^{2q}}$$

When the cell proportions $\pi_1 = \pi_2 = \cdots = \pi_\Lambda = 1/\Lambda$ within the original *m*-way table T[U(*n*)], higher values of *q* yielded lower minimum total approximate probabilities of successful disclosures. It was observed that when the cell proportions were fairly balanced among the Λ total cells in $T[U(n)]$ (that is, no one cell in $T[U(n)]$ contains a very large proportion of counts relative to its remaining Λ-1 cells), higher values of q yielded lower probabilities of successful disclosures from the differencing attack $T[U(n-q)] - T[U(n-1-q)] = T[U(1)]$. For example, Figure 4.3.6 shows a two-way table for gender by race $T_1[U(n)]$, where the cells proportions in $T_1[U(n)]$ are fairly balanced among its 8 total cells. Therefore, for any given pair of independently subsampled tables $T_1[U(n-q)]$ and $T_1[U(n-1-q)]$, the differencing attack $T_1[U(n-q)] - T_1[U(n-1-q)] = T_1[U(1)]$ will yield smaller approximate probabilities of successful disclosures for *gender* and *race* for larger values of q, as shown in Table 4.3.6.

However, it was also observed that if one cell in $T[U(n)]$ contains an very high proportion of counts relative to its remaining Λ-1 cells, then the approximate probability of obtaining a successful disclosure from $T[U(n-q)] - T[U(n-1-q)] = T[U(1)]$ remained high, regardless of the value of q. For example, Figure 4.3.7 shows a two-way table for gender by race, $T_2[U(n)]$, where the cell for *gender* = female and *race* = white contains a very high proportion of counts, 0.9814, relative to its remaining 7 cells. Therefore, for any given pair of independently subsampled tables $T_2[U(n-q)]$ and $T_2[U(n-1-q)]$, the differencing attack $T_2[U(n-q)] - T_2[U(n-1-q)] = T_2[U(1)]$ will still yield high approximate probabilities of successful disclosures for *gender* and *race*, even for higher values of q, as shown in Table 4.3.7. It is important to note that the sensitive cell (the cell that differs by only one count in both $T[U(n)]$ and $T[U(n-1)]$, as shown in Figure 4.2.1) does not need to be the cell that contains the highest proportion of counts relative to the remaining Λ-1 cells within the original table $T[U(n)]$. If any cell within T[U(n)] contains a very high proportion of counts, the approximate probability of obtaining a successful disclosure from the differencing attack $T[U(n-q)] - T[U(n-1-q)] = T[U(1)]$ will still remain high.

$T_1[U(n)]$	*race*			
gender	White	Black	Asian	Other
Male	0.123	0.111	0.145	0.152
Female	0.116	0.133	0.101	0.119

Figure 4.3.6

$T_2[U(n)]$	*race*			
gender	White	Black	Asian	Other
Male	0.0016	0.0017	0.0088	0.0007
Female	0.9814	0.002	0.0015	0.0023

Figure 4.3.7

q	Total Approximate Probability of Obtaining a Successful Disclosure from $T_1[U(n-q)] - T_1[U(n-1-q)] = T_1[U(1)]$, for the Given Cell Probabilities in $T_1[U(n)]$
$q = 2$	0.03014660
$q = 3$	0.01000567
$q = 4$	0.00412943
$q = 6$	0.00107119
$q = 8$	0.00039571
$q = 10$	0.00018161
$q = 15$	0.00004419
$q = 20$	0.00001626

Table 4.3.6

q	Total Approximate Probability of Obtaining a Successful Disclosure from $T_2[U(n-q)] - T_2[U(n-1-q)] = T_2[U(1)]$, for the Given Cell Probabilities in $T_2[U(n)]$
$q = 2$	0.9280158
$q = 3$	0.8942550
$q = 4$	0.8618921
$q = 6$	0.8011097
$q = 8$	0.7451983
$q = 10$	0.6937315
$q = 15$	0.5820591
$q = 20$	0.4907147

Table 4.3.7

To increase the confidentiality protection against differencing attacks, we are exploring possible modifications to the current *drop q* subsampling routine.

5. Recent Work on the Microdata Analysis System: The Cutpoint Program

5.1 The Cutpoint Generation Program

Cutpoints are used to create buckets or bins of numeric values. We require that each bin contains at least a pre-specified number of observations (*d*), otherwise, a universe defined from one of these bins may fail the *No Marginal 1 or 2* or *75 Rules*. It is common for a bin to represent a range of numeric values, e.g., ages 20 to 25. Doing this adds an additional level of uncertainty since all the analyses will be based on a universe defined on ranges of values instead of single value. This also adds further protection against a differencing attack, since forming two similar universes by incrementing the cutpoint value by one may result in an imprecise difference.

There are a number of approaches for generating cutpoints. The ones we consider are the following: fixed width bins, minimum width bins, increasing width bins, and partitioned bins. The fixed width bin approach ensures that the width of each bin is the same. In other words, the difference between the maximum bin cutpoint value and the minimum bin cutpoint value is the same for every bin. The minimum width bin approach creates bins with as close to *d* observations in each bin. These bins vary in size. The bin widths tend to be smaller than the other approaches, leading to bins of a finer granularity. The increasing width bins approach gradually increases the width of the bins. Based on Steele and Zayatz (2006), this approach begins with a fixed bin width that increases as numeric values increase. For example, the bin width *d* may equal 50 when the numeric variable values are less than 200, but increase to 100 once variable values get larger. Finally, unlike the other methods, partitioned binning uses a top down strategy for bin generation. Beginning with the entire set of values, this strategy recursively partitions the sorted data until there are approximately *d* observations in each bin. Using this approach, bin widths are not equal, but are multiples of each other.

Each of these approaches has a number of strengths and weaknesses depending upon the range and distribution of the variable in question. At this stage, the partitioned binning seems the most promising. However, we are considering hybrid strategies that use different approaches based on the statistical properties of a given variable.

6. Future Work

For future work, we will continue to develop a beta prototype of the Microdata Analysis System as part of DataFERRETT (Chaudhry, 2007). We will begin to test both the software itself and the confidentiality rules implemented in the MAS beta prototype to test their effectiveness in preventing the disclosure of confidential data. In addition, we will be adding more data sets and more types of statistical analyses to the system.

References

Chaudhry, M. (2007). “Overview of the Microdata Analysis System,” *Statistical Research Division Internal Report*, U.S. Census Bureau.

Duncan, G. T., Keller-McNulty, S., and Stokes, S. L. (2003). “Disclosure Risk vs. Data Utility: The R-U Confidentiality Map,” *Technical Report 2003-6*, Heinz School of Public Policy and Management, Carnegie Mellon University.

Gomatam, S., Karr, A.F., Reiter, J.P., Sanil, A. (2005). “Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Servers,” *Statistical Science*, 20, 163-177.

Kaufman, S., Seastrom, M., and Roey, S. (2005). “Do Disclosure Controls to Protect Confidentiality Degrade the Quality of the Data?,” *Proceedings of the Section on Survey Research*, American Statistical Association.

Lucero[1], J. (2009). “Evaluation of the Effectiveness of the *Drop q Rule* Against Differencing Attack Disclosures,” *Statistical Research Division Confidential Research Report Series* #????, U.S. Census Bureau (In Progress).

Lucero[2], J. (2009). “Confidentiality Rules for Universe Formation and Geographies for the Microdata Analysis System,” *Statistical Research Division Confidential Research Report Series* #????, U.S. Census Bureau (In Progress).

Lucero[3], J. (2009). “Confidentiality Rule Specifications for Performing Regression Analysis on the Microdata Analysis System,” *Statistical Research Division Confidential Research Report Series* #????, U.S. Census Bureau (In Progress).

Reiter, J.P. (2003). “Model Diagnostics for Remote-Access Regression Servers,” *Statistics and Computing*, 13, pp. 371-380.

Reiter, J.P. (2004). “New Approaches to Data Dissemination: A Glimpse into the Future?,” *Chance*, 17:3 (Summer 2004), 12-16.

Reznek, A.P. (2003). “Disclosure Risks in Cross Section Regression Models,” *Proceedings of the American Statistical Association, Government Statistics Section*, [CD-ROM], Alexandria, VA, American Statistical Association.

Reznek, A.P. and Riggs, T.L. (2004). “Disclosure Risks in Regression Models: Some Further Results,” *Proceedings of the American Statistical Association, Government Statistics Section*, [CD-ROM], Alexandria, VA, American Statistical Association.

Rowland, S. and Zayatz, L. (2001). “Automating Access with Confidentiality Protection: The American FactFinder,” *Proceedings of the Section on Government Statistics*, American Statistical Association.

Steel, P. and Zayatz, L. (2006). “Description of a Microdata Access System” for Presentation to the Census Advisory Committee of Professional Associations, US Census Bureau, October 27, 2006.

Weinberg, D., Abowd, J., Rowland, S., Steel, P., and Zayatz, L. (2007). “Access Methods for United States Microdata,” *Center for Economic Studies* Paper No. CES-WP-07-25, U.S. Census Bureau.

Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control, Springer-Verlag New York, Inc.