NISS

National Institute of Statistical Sciences
PO Box 14006, Research Triangle Park, NC 27709–4006
Tel: 919.685.9300 FAX: 919.685.9310
www.niss.org

National Center for Health Statistics...*Monitoring the Nation's Health*

Data Confidentiality: The Next Five Years

May 1–2, 2008, Hyattsville, MD

Co-Sponsors
Office of Research and Methodology, National Center for Health Statistics, CDC
National Institute of Statistical Sciences

Program

Thursday, May 1, 2008

9:30 AM *Welcome and Introductions*

10:00 *Session I: Query Systems*
- Cynthia Dwork, Microsoft Research: Differential Privacy: What we Know and What We Want to Learn
- Adam Smith, Pennsylvania State University: Integrating Differential Privacy with Statistical Theory
- Chair: William Winkler, Census Bureau)

11:45 Lunch on your own

1:00 PM *Session II: Weighted Data*
- Avinash Singh, Carleton University: Maintaining Analytic Utility While Protecting Confidentiality of Survey Weighted Data
- Stephen Fienberg, Carnegie Mellon University: The Relevance or Irrelevance of Weight for Confidentiality and Statistical Analyses
- Chair: Lawrence Cox, NCHS

2:45 Break

3:15 *Session III: Distributed Data*
- Alan Karr, NISS: Secure Statistical Analysis of Distributed Data
- Xiadong Lin, University of Cincinnati: Privacy-Preserving Distributed Maximum Likelihood Estimation
- Chair: Rebecca Wright, Rutgers University

5:00 Adjourn for the Day

Friday, May 2, 2008

8:30 AM *Session IV: Synthetic Data*
John Abowd, Cornell University: Synthetic Data and Randomized Sanitizers
Jerome Reiter, Duke University: Some Next Steps in Synthetic Data Research
Chair: Laura Zayatz, Census Bureau

10:15 Break

10:45 *Session V: Tabular Data*
Lawrence Cox, NCHS/CDC: A Data Quality and Confidentiality Assessment of Complementary Cell Suppression
Aleksandra Slavkovic, Pennsylvania State University: Tabular Data: From Margins to Margins and Conditionals
Chair: Joe Fred Gonzalez, Jr., NCHS

12:30 PM Lunch on your own

2:00 *Panel Discussion: Federal Agency Needs*
Jacob Bournazian, EIA
Lawrence Ernst, BLS
Marilyn Seastrom, NCES
Chair: Alan Karr, NISS

3:30 Workshop Adjourns

Abstracts

Differential Privacy: What we Know and What we Want to Learn

Cynthia Dwork
Microsoft Research

The problem of statistical disclosure control—revealing accurate statistics about a population while preserving the privacy of individuals—has a venerable history. An extensive literature spans multiple disciplines: statistics, theoretical computer science, security, and databases.

This talk surveys a body of work revisiting the problem from a cryptographic perspective. We describe an "ad omnia" (as opposed to "ad hoc") definition of privacy, called differential privacy and show at least one way of achieving it. We then discuss the merits of this approach, using as an example a differentially private method for contingency table release, in which a user may request an arbitrary set of tables and consistency among released tables may be required.

Finally, we turn to what we do not understand, and outline what we feel are the most pressing questions.

Integrating Differential Privacy with Statistical Theory

Adam Smith
Pennsylvania State University

I'll discuss why the seemingly stringent requirements of differential privacy (discussed in Cynthia Dwork's talk) are compatible with important elements of statistical theory. Specifically, for several important types of estimators, there exist private analogues with same asymptotic error and computational cost:

1. For any sufficiently regular parametric model, there exists a differentially private, asymptotically efficient point estimator for the parameters, i.e. an estimator with the same convergence rate as the maximum likelihood estimator (MLE);
2. For any smooth density on the interval [0, 1], there exists a private histogram estimator whose expected L_2 error is the same as that of the optimal, non-private fixed-width histogram estimator.

These results suggest that one can get the benefits of rigorously defined privacy guarantees while preserving statistical validity. Many open questions remain, however, some of which I will try to highlight in the talk.

Maintaining Analytic Quality while Protecting Confidentiality of Survey Data

Avinash C. Singh
Carleton University

We consider the problem of providing a public use file from a data base obtained from a sample survey such that disclosure of individuals via indirect identifiers is reasonably well protected while quality of statistical analysis after disclosure treatment is reasonably well maintained. We address the common concern of respondents regarding the inside intrusion scenario where the intruder might know the presence of the target in the database. This implies that some form of treatment via random perturbation (i.e., substitution of key identifying variables-IVs) and random suppression (i.e., sampling-out of individual records) is needed to introduce uncertainty about the identity and presence of the target.

With survey data, information about stratum and PSU (primary sampling unit or cluster) subsets of the data is needed for variance estimation in data analysis, although it is sufficient to have only pseudo identifiers for these data subsets. However, this information may provide additional IVs for the intruder to narrow down his search. For this problem, it turns out that it may not be necessary to further randomly treat these additional IVs because of the treatment at the lower level (i.e., at the individual) which induces uncertainty about the presence and identity of the individual. Moreover, with survey data, sampling weights are present in the data base which are important for avoiding selection bias in analysis with nonignorable designs. Now, the original sampling weights reflecting inverse of selection probabilities may act as IVs if the intruder has some information about the design. However, typically the sampling weights are calibrated to adjust for nonresponse, poststratification , and extreme values which diminishes considerably their value as IVs. The problem is further mitigated if a second stage of calibration is performed after the disclosure treatment phase so that estimates for a selected set of analytic variables do not change after the treatment.

With known random mechanisms for disclosure treatment via substitution and subsampling, it is possible to compute measures of disclosure risk at various levels: the individual, subgroup, and the whole data base, under mild nonparametric assumptions. At the record level in the treated data base, for an individual appearing to be the target, these measures are computed from probabilities of surviving the treatment, being correctly or incorrectly classified as uniques or nonuniques, and then having values of the sensitive variables that puts them at risk. Here, the second set of selection probabilities introduced in the treatment phase could be used to estimate appropriate risk parameters defined for subgroups of records in the original data base.

The analytic quality measures for the treated database in terms of RRMSE (relative root mean square error) and RB (relative bias) for means and totals corresponding to selected analytic variables can be computed using survey sampling methods by regarding the original data base as the population, and the disclosure-treated one as the sample. Similarly, estimates of the model parameters after treatment can be compared with the original estimates. Here the variance of the estimates would need to be adjusted for substitution analogous to imputation adjusted variance estimation in survey sampling. Moreover, in making inference about the finite or super population parameters using the treated data set, two phase sampling techniques can be used because the original data base itself represents the first phase of sampling. If the original data base is obtained from census or administrative sources, then the simpler single phase sampling techniques would naturally be applicable to the treated data set.

The above considerations suggest in a natural way that survey sampling techniques can be employed for our purpose in protecting confidentiality and quality of data. In fact, there is a strong analogy between taking a census that provides full information about the whole population, and releasing an untreated complete data base. To avoid huge monetary cost, well designed sample surveys are conducted at the price of introducing error in the estimates due to under/over coverage and sampling variability, but this is controlled via sample allocation under constraints on precision and bias of selected estimates. Moreover, sample estimates are adjusted via calibration methods to deal with possible bias due to nonresponse and coverage error as well as instability due to extreme weights. Similarly, to avoid huge disclosure cost (tangible and intangible), the data can be stratified into risk strata or micro agglomerates based on uniqueness criterion, and then treatment rates for random selection of records for substitution and subsampling can be allocated to provide control on bias due to substitution and variance due to subsampling in key set of estimates. Thus, one can have a simultaneous control on disclosure risk and analytic utility which tend to work against each other as was originally formulated by Duncan and Lambert. With this motivation in mind, a method termed MASSC (signifying micro agglomeration with substitution, subsampling, and calibration) was developed by Singh (2002, 2006, www.uspto.gov/patft/index.html); see also Singh, Yu, and Wilson (2004, ASA SRMS Proc.).

A review of the above background along with steps needed to implement MASSC will be presented and illustrated with a simple example. Relative merits and demerits of MASSC with alternative methods such as those suitable for tabular data of counts and magnitudes, and synthetic methods will be discussed. In this context, some important contributions are, among others, due to de Waal and Willenborg (1997), Trottini and Fienberg (2002), Skinner and Carter (2003), Raghunathan, Reiter, and Rubin (2003), Cox, Kelly, and Patil (2004), Winkler (2004), and Reiter (2005). Finally, it may be remarked that MASSC produces a nonsynthetic treated data set, and is applicable to any data set (macro or micro) that can be represented as a rectangular file with rows corresponding to individuals and columns to identifying, sensitive, and other analytic variables.

Secure Statistical Analysis of Distributed Data

Alan Karr
NISS

Secure multiparty computation protocols from computer science enable principled, secure analysis of horizontally partitioned, distributed databases when the sufficient statistics for the analysis are additive across database owners. When this structure is present, in fact only secure summation is required.

Following a brief introduction to the secure analysis for vertically as well as horizontally partitioned data, the talk will focus on a number of unresolved—and in most cases, not even well-understood issues, which include data pre-processing, allowing analysts to be good statisticians, protections against dishonesty, risks arising from unequal database sizes and data heterogeneities, non-additivity, methods that allow agencies to opt out based on the results of the analysis, and numerical and algorithmic issues.

Privacy-Preserving Distributed Maximum Likelihood Estimation

Xiadong Lin
University of Cincinnati

Although statistical analysis over combined data possesses huge potentials in knowledge discovery, it can also induce great disclosure risks. Thus, there is a need to develop statistical tools that can obtain proper analysis results while preserving data privacy. Individual privacy preserving statistical analysis protocols have been proposed for specific statistical models in the past few years. In this talk, I will present methods and protocols for privacy preserving maximum likelihood estimation in general settings. These methods can be used in various statistical models that utilize maximum likelihood for parameter estimation. I will discuss models and solutions for both the horizontally and vertically partitioned data, and proposed procedures that give participating parties the choice to withdraw from joint computations.

Synthetic Data and Randomized Sanitizers

John Abowd
Cornell University

Synthetic data were originally proposed as a statistical disclosure limitation method that protected confidential data by releasing only statistically valid pseudo-data rather than the underlying data themselves. Statistical validity was achieved by sampling from the estimated posterior predictive distribution of the underlying data. While the statistical validity of synthetic data was comparatively easy to define, suitable definitions of the confidentiality protection it affords have proven more elusive. A random sanitizer is any function that maps data and noise into a response to a set of queries. Synthetic data are one of many random sanitizers that can be applied to confidential data. Given a specific set of criteria for random sanitizers, formal definitions of confidentiality protection, and specific analytical validity objectives, one can evaluate synthesizers for their analytical validity and disclosure risk trade-off. A complete research program would integrate all random sanitizers and a variety of analytical validity measures into this framework, with the goal of providing statistical disclosure limitation criteria that can be understood and used by the custodians of publicly collected data.

Some Next Steps in Synthetic Data Research

Jerome Reiter
Duke University

In this talk, I discuss some important challenges to effective and widespread use of synthetic data methods for public use data. These include methods for generating synthetic data, methods for providing automated feedback on the validity of analyses conducted with synthetic data, and methods for selecting which genuine values to synthesize. I describe some broad principles for approaching each challenge.

A Data Quality and Confidentiality Assessment of Complementary Cell Suppression

Lawrence Cox
NCHS/CDC

Complementary cell suppression has been used to control statistical disclosure in tabular data by many statistical organizations over several decades. Its use has been proliferated by software developed, most notably, by the Census Bureau, Statistics Canada, the CASC Project (European Union), and the National Center for Health Statistics. Suppression impacts data quality adversely by removing both sensitive data and nonsensitive data (complementary suppressions). It has been suggested that these effects can be mitigated by publishing exact interval estimates of suppressed cells (which are already available to the sophisticated user via linear programming, albeit at considerable effort). The principal disclosure rules driving cell suppression are the p-percent and p/q-ambiguity rules. We examine the effects of cell suppression and interval estimates on data quality and the extent to which exact interval estimates can be used to reduce data protection under p-percent and p/q-ambiguity rules.

Tabular Data: From Margins to Margins and Conditionals

Aleksandra Slavkovic
Pennsylvania State University

Work on statistical methods for confidentiality and disclosure limitation have seen coupling of tools from statistical methodologies and operations research. For the summary and release of data in the form of a contingency table the methodology has primarily focused on evaluation of bounds on cell entries in k-way contingency tables given the sets of marginal totals, with less focus on evaluation of disclosure risk of other summaries such as conditional probabilities, that is, tables of rates. Narrow intervals—especially for cells with low counts—could pose a privacy risk. We present the closed- form solutions for the linear relaxation bounds on cell counts of a contingency table given full and partial conditional probabilities thus significantly minimizing the need for a computing time. We also compute the corresponding sharp integer bounds via integer programming and show that there can be large differences in the width of these bounds, suggesting that using the linear relaxation is often an unacceptable shortcut to estimating the sharp bounds and the disclosure risk for the tables of rates. However, for large sparse contingency tables this is prohibitively time-consuming for most practical usages, and thus the closed-form linear relaxation bounds could be a useful easy estimate for an agency deciding whether to release such summaries. We also discuss how the sharp bounds given partial conditional information relate to the sharp bounds given corresponding marginals.

We will also briefly discuss the effects of releasing the conditional rates on data utility using various distortion measures and log-linear and logistic regression models. This work is tied to generation of "synthetic data," that is tables that preserve certain margins and/or conditionals, by using tools from computational commutative algebra such as Markov bases.