Data Access and Personal Privacy: Appropriate Methods of Disclosure Control

A statement by the American Statistical Association.

Approved December 6, 2008

Statement:

Access to high quality data is essential to advancing science and improving the human condition. Robust new sources of data on human behavior allow researchers to ask and answer complex questions and hence guide policy decisions. Powerful and sophisticated electronic technologies have made much of this data readily accessible to the public.

At the same time, much of this data contains personal information, so these electronic tools for combining and analyzing publicly accessible data pose a distinct threat - in perception if not in reality -- to privacy, as well as a potential for inflicting great harm on persons and establishments. The protection of personal privacy is of paramount importance in engaging the cooperation of respondents, and thus in producing and distributing the high quality data needed for research. Fortunately, modern statistical tools have been developed to help ensure the appropriate treatment of confidential information while still making useful data available for public policy and scientific advancement.

This statement is intended to provide the American Statistical Association's (ASA) perspective on the assessment of the risk associated with data dissemination and an overview of the way in which statisticians can help limit that risk.

The ASA urges distributors and users of data, particularly sensitive data such as public health and biologic data, to familiarize themselves with risk assessment, and to consult with statistical professionals when necessary. The ASA further urges the media to be mindful of these issues when it presents data to the public.

Context:

Many forms of data are collected and disseminated to guide both research and policy decisions. For example, health data on individuals are collected and used by state agencies and others so that trends can be monitored, potential public health hazards can be identified, and public health can be protected. At the same time, the privacy of the individuals who provide the data must be safeguarded. An illustration of the tension between these sometimes conflicting needs occurred in Delaware in 2008. The Delaware press sought detailed information about the location and characteristics of certain cancers. However, the Delaware Division of Public Health cited privacy concerns in refusing the release of such data. In response to these concerns, the state passed legislation (Delaware Senate Bill 235, now state law) requiring the release of such data but allowing the agency to take steps first to protect patient privacy.

In the first instance, statisticians can follow well established fair information practices to protect privacy, such as collecting only the information that is needed, articulating the purpose of the information collection, and providing informed consent. A critical element of informed consent is to accurately explain what assurances of confidentiality are available.

Statisticians also have a long history of studying ways to protect the confidentiality of data while providing information to policymakers. The traditional way of ensuring confidentiality while disseminating data has been to aggregate information and report it in tables. This approach generally acts to mask information that might specifically identify anyone.

The challenge of safeguarding confidentiality has become more difficult for data custodians. Many new forms of data on human behavior, such as video data, biologic samples, or transaction data, are not particularly useful to researchers or policy makers in tabular form. As a result, such "micro-data" is often disseminated after the information is "de-identified." Unfortunately, statistical research shows that such de-identification is often insufficient and could result in a breach of confidentiality if reidentification were attempted by an individual with the right skills, a computer, and access to publicly available databases.

In one example, a student at the Massachusetts Institute of Technology showed that 97 percent of the names and addresses on the 1997 voting list for Cambridge, Massachusetts were unique using only zip code and date of birth[1]. The same research showed that this same information, along with medical insurance claims records of state employees, was contained in files made available to researchers by the Massachusetts Group Insurance Commission. By comparing the two sources, the records of the Governor of the state were re-identified, even though his personal identifiers had been removed from the insurance records.
In another example, geneticists who have made substantial progress in the mapping of the human genome also have found that there is reason for caution in making genetic information generally available[2]. The increased availability of genomic data for research, coupled with demonstrations that conventional protective procedures do not completely mask the presence of an individual's genetic material in certain databases, has led to measures for increased security.

Today we are developing better statistical tools that can help guide the proper release of data. First, those tools can help ensure the proper assessment of risk. Second, the tools help ensure the proper treatment of confidential information, so that confidential facts do not become public knowledge through the apparently harmless release of aggregated data or de-identified micro-data. Statisticians, working with computer scientists and others, can help ensure continued access to research data while protecting the privacy of the individuals from whom the data came.

A brief discussion of the statistical resources available follows. For further information, please contact the chair of the ASA's Privacy and Confidentiality Committee. This contact information can be obtained from the committee's website, or by calling the American Statistical Association, 703-684-1221.

Background

The ASA recognizes that risk assessment and confidentiality protection are not simple matters. It believes that statistical techniques are essential to identifying and preventing potential disclosures and invaluable to resolving them. *However routine or unusual the information to be protected, the statistician can considerably enhance its usefulness while also protecting privacy*. The ASA emphasizes the following points:

Confidentiality protection is important

The protection of personal privacy is of paramount importance in the production and distribution of statistical data.

The quality of those data is strongly influenced by the public's trust that pledges of confidentiality will be rigorously observed.

Data access is important

Optimizing access to high quality data is critical to informed decision making and is the principal justification for their collection and dissemination.

The assurance of confidentiality is a primary concern in considering what scope and extent of access to personal information will be granted.

Statisticians can play an important role in ensuring that both goals are met, and need to work with data users, data producers, and data custodians to accomplish these goals.

The sharing and dissemination of information gathered under a pledge of confidentiality must be subject to rigorous statistical scrutiny to ensure consistency with the confidentiality pledges.

The profession of statistics has developed the requisite tools to help with the appropriate treatment of confidential information. Additionally, the profession is actively engaged in research to further refine these tools and to develop means to make useful information available for public policy and scientific advancement.

The Assessment of Risk in Statistical Data

The assessment of risk depends on the way in which the information is produced. Until fairly recently, the production of information for dissemination to the public relied principally on printed, tabular data. Statisticians have long been sensitive, therefore, to the potential risk of disclosure in such data. Although tables are intended to protect individual information by presenting grouped figures, there are situations in which the size and/or the distribution of those groups can reveal more information about individuals or business establishments than had been publicly known.

In contrast to tabular data, which are presented in aggregate form, the information contained in micro-data is disaggregated. The information contained is specific to the individual. An electronic micro-data file may contain many thousands of data records, each referring to a separate person. This format permits the researcher to specify with exactitude the kind of questions that can be addressed and to utilize much more powerful analytic tools. This very advantage, however, carries with it the possibility of identifying one or more respondents - and the more detailed the information, the more individual records become distinct from each other, making study participants easier to identify.

It is a common misconception that once names and certain other direct identifiers (address, telephone number) that lead directly to a person have been removed from a body of information, the remaining data may be judged safe and can be shared without risk of compromising the privacy of the data providers.

Statisticians and members of other professions, however, have demonstrated repeatedly that modern computational technology and the widespread availability of personal information on the Internet can render information quite as revealing as though the names had not been removed. That is, a seemingly anonymous body of data could be rendered identifiable.

Techniques for Protecting Confidentiality

Taking the aforementioned facts into account, statistical scientists and agencies responsible for developing and distributing data have developed a variety of counter measures to *de-identify* statistical databases to block efforts to manipulate them to disclose personal information. Strategies for preventing unauthorized and inappropriate disclosure of identifiable information generally involve some combination of modification of data content and restriction of data access. The first strategy involves some loss of information detail and the second, while permitting access to more complete data to qualified users, limits who, under what conditions, and for what purpose they may be used. Thus the selection of a strategy involves a careful consideration of the interests of legitimate data users while strictly adhering to confidentiality protections promised to the subjects of the data; by selecting among a variety of strategies a satisfactory resolution can often be found. Alternatives that have been considered include:

- Modifying the values of information items to maintain statistical quality but avoid disclosures. One such strategy is to blur or disguise the data in such a way that individual data items cannot be uniquely associated with or attributed to a particular person or establishment.
- Distributing synthetic data sets whose variables have the same statistical distributions and relationships as the original data from which they are derived but containing no actual information from the original data. Partially synthetic files are another way to avoid disclosures while keeping the bulk of the data intact.
- Providing access to detailed data only in restricted data enclaves where appropriateness of use can be monitored, access is restricted to authorized individuals, and those individuals are trained in confidentiality protection. The enclaves can be set up either to require the analyst to be physically present at the restricted site or to allow remote access to authorized analyses.
- Permitting tailored online data analysis of detailed databases with results subjected to disclosure avoidance review.
- Making selected information files available under licensing arrangements that guarantee secure and confidential handling of data by trusted researchers.

Various Federal agencies' data access policies employ one or more modes of access, and some are able to provide different tiers of access, for example providing a minimally protected dataset (e.g., no direct identifiers like names) in secure enclaves, slightly restricted data (e.g., with aggregated geography) via licensing, and blurred or synthetic data for unrestricted public use.

A comprehensive review of disclosure avoidance techniques is beyond the scope of this statement, but the reader is referred to the Privacy, Confidentiality and Data Security website[3] for a convenient source of references on current regulations, recommendations, and best practices in the field. Three recent publications deserve special mention: the *Report on Statistical Disclosure Limitation Methodology*[4] issued by the Federal Committee on Statistical Methodology, the *Handbook on Statistical Disclosure Control*, a product of the Centre of Excellence in Statistical Control[5], and *Expanding Access to Research Data: Reconciling Risks and Opportunities*, a report of the National Academy of Sciences[6]. Together these publications represent an up-to-date perspective based on a vast amount of experience and expertise.

Appendix: Definitions

Key to the understanding of disclosure avoidance are the concepts of privacy, confidentiality, and data protection[7]. Informational privacy encompasses an individual's freedom from excessive intrusion and the ability to choose the extent and circumstances under which one's personal information will be shared with or withheld from others. The assurances given to information providers concerning the care and potential sharing of this information are detailed in a pledge of confidentiality. Data protection refers to the set of policies and procedures that ensure that the protection promised is actually provided. These policies and procedures are generally quite comprehensive and involve administrative, physical, and electronic safeguards. When information is shared with the public or with parties not included in the pledge of confidentiality, the possibility of disclosure arises. It is at this point that the statistician's skills and experience come into play.

Information may be shared as *tabular* or *micro-data*. The former is represented by data grouped according to one or more characteristics. A simple table would contain categories (cells) of age and gender, and more complex tables contain additional classifications (e.g. race, income, etc.). Each one of these more complex tables could be constructed for a number of other variables or dimensions (e.g. a separate complex table for each state in which data were collected). The table entries may be the actual number of respondents falling into a given category (frequencies) or an average, rate, percentage or other quantity that applies to respondents so categorized

[1] Sweeney, L. Computational Disclosure Control: A Primer on Data Privacy Protection. Doctoral Dissertation, Massachusetts Institute of Technology, May, 2001.

[2] NIH Background Fact Sheet, August 28, 2008. **http://grants.nih.gov/grants/gwas/**. See also, Homer N, Szelinger S, Redman M, et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genet 4(8): e1000167. doi:10.1371/journal.pgen.1000167

[3] The PCDS website is found at **http://www.amstat.org/committees/pc/index.html**.

[4] Federal Committee on Statistical Methodology, Working Papers Nos. 22 (Second version, 2005). Found at **http://www.fcsm.gov/reports/#fcsm**.

[5] Center of Excellence for Statistical Control, Handbook on Statistical Disclosure Control, Version 1.01, March 2007. Found at **http://neon.vb.cbs.nl/cenex/**.

[6] National Academy of Sciences, *Expanding Access to Research Data: Reconciling Risks and Opportunities, 2005*.

[7] For a more detailed discussion of these and related terms, see **http://www.amstat.org/committees/pc/keyterms.html**.