between privacy and access concerns, while maintaining high levels of participation and truthfulness in surveys—without public support.

I believe the time has arrived for the statistical system to "go public" with confidentiality and data access issues. This is based on my opinion that both the policy debate and the development of technical solutions to disclosure protection problems within the statistical system have matured sufficiently to be analyzed and discussed at a general level. This was not the case 10 years ago. There are dangers to raising these issues, however, that must be kept in mind: By raising the issue, it can be made salient in a way that frightens the public; also, we run the risk of confusing people with arcane, inconclusive or contradictory technical and legal information (thereby eroding their confidence in a different way).

The approach should be to communicate the importance of reliable and varied statistics to the society and the economy, while instilling confidence that individual respondents have rights, the protection of which is the bedrock of the statistical system. The approach should first be made through influence groups: advocates for privacy; groups representing the disadvantaged (e.g., the homeless) or those at risk (e.g., AIDS); those concerned with the rights of individuals (e.g., ACLU); the press; and those concerned with the political process. Avenues to many of these groups exist already within the normal workings of the statistical system. Well- and nontechnically articulated arguments need to be developed and discussed with these groups, leading perhaps to experiments of one kind or another, and, ultimately, consensus and change. If these groups are convinced and, to a degree, become advocates for the statistical system on issues of privacy and data access, I believe the support of the public at large will follow.

Within these deliberations, it is important to maintain a focus on legislative issues. Laws do not prescribe how statistical agencies are to design questionnaires and samples, estimate parameters or edit questionnaires and impute missing or faulty data; yet regarding confidentiality, many laws are absolute, one-side in assigning penalties and, although written in the absence of technical information, exert a driving influence on agencies' confidentiality practices. Most agree that responsibility for disclosure protection should, like the data, be shared between the data provider and the data user. This strikes me as an issue easily understood and potentially supportable by outside groups.

The Duncan–Pearson article does a good job of presenting the mounting issues faced by the statistical system along the data confidentiality/data access front. It is readable outside the statistical community, and that is important if we are to broaden the discussion, as I suggest be done.

Comment

Sallie Keller-McNulty

I would like to commend Duncan and Pearson for their contribution on the very important topic of data access and confidentiality. I am pleased that *Statistical Science* has had the foresight to publish such an article, and I hope that many researchers will read and react to the material. I have no disagreements with the opinions expressed in this manuscript, but I would like to bring more attention to a few points that were made.

First, I would like to comment on the various ways that disclosure has been conceptualized. In particular, attention has been focused on inferring *attribute* values. I contend that, with the database systems as a data storage and access medium, we need to also be concerned with the direct disclosure of *relationships* between attributes. We have been conditioned to view data as a file or rectangular array where the columns represent attributes, the rows represent data records (one for each respondent), and the entries within a row represent attribute values. Attribute disclosure is conceptualized as inferring an element of this array. In this setup, hypotheses about the relationships between attributes are validated through analysis of the data. In a database, relationships among attributes are contained in the schema, or logical structure, of the system. Relationships as well as attribute values are considered objects (i.e., encapsulation of values with their semantic meanings). Disclosure in a database system can be defined as inferring an

Sallie Keller-McNulty is Associate Professor of Statistics, Kansas State University, Manhattan, Kansas 66506.

object. This implies that relationships as well as attribute values could be compromised directly in this environment. It could be argued that this is a trivial point, because in the universal form of a database all the attributes of the database exist in every record and the juxtapositions of these attributes represent the information contained in the schema. Compromising an attribute and its location within the universal form could constitute learning a relationship. However, the universal form is a theoretical model with many records per respondent. A database system is stored in a compressed form with the relationships explicitly included, thus at risk.

Duncan and Pearson's characterization of matrix masking a microdata file as $M = AXB + C$ is excellent. This characterization will also apply to masking a database where X is the universal form of the database system. Research will be needed to determine how to map A, B and C to the objects of the database. In this context, I would like to clarify one possible point of confusion for the reader. Throughout their article, Duncan and Pearson refer to disclosing Y from X. In the situation of a matrix mask, X is M and Y is frequently X.

Duncan and Pearson touch briefly on the important issue of the need for new statistical procedures to analyze data with sophisticated masks and the need to train researchers in the use of these procedures. This leads to the topic of cost. There are serious tangible and intangible costs associated with the problem of data access and confidentiality. The tangible costs are implementation costs, access costs and educational costs. The intangible costs are a function of the accuracy and usefulness of the released data. As methods are developed for the protection of data, their tangible and intangible costs should also be analyzed because policy decisions on data security will clearly be a function of cost.

Finally, I would like to comment on Duncan and Pearson's discussion of *informed consent*. The examples they cite deal with highly educated respondents. Focusing any data confidentiality policies on informed consent for the general populous may be overly optimistic. The methods used to compromise data are complex, and our ability to measure disclosure risk and conceptually model the decision process of the data spy is limited. It is probably unrealistic to believe that we can convey the degree of risk for a respondent's data through an informed consent agreement. While informed consent agreements may hold up in court, I do not believe they are an ethical solution to the data confidentiality problem.

The focus in Duncan and Pearson's article is on federally collected data. The dilemma between data access and confidentiality exists in privately collected data as well. Computer scientists have been concerned about data security in the private sector for quite some time. Much of their research approaches the problem by considering sophisticated database management systems which will store and release the data (Adam and Wortman, 1989; Keller-McNulty and Unger, 1991). As Duncan and Pearson point out, with new technology and easier access to information, our federal agencies will soon be confronted with managing data in complex database system. A concrete example is to view our nation's public data as a decentralized distributed database. The links in the data from one agency to the next exist but have simply not been connected electronically. Reviewing approaches computer scientists have taken to handle security in distributed database management systems should give us insight into what policies might be valuable in governing the access and security for sharing data across federal agencies. In general, I suggest that we as statisticians follow closely the progress in computer science on data and computer security. Through collaboration with computer scientists, we might avoid serious duplication of effort toward a solution to the complex problem of providing rich data yet maintaining adequate confidentiality.