DISCUSSION

Diane Lambert, Carnegie-Mellon University[1]

The goal of releasing data that are useful for making accurate inferences about a group conflicts with the goal of protecting attributes of individuals belonging to the group. For example, an accurate evaluation of CETA programs requires geographical information at the microdata level to account for local differences in welfare programs and employment opportunities (cf. Boruch and Cecil [2]), but release of detailed geographical information increases the risk of disclosure of sensitive information about individuals.

There are two kinds of resolutions of the disclosure conflict: administrative rules to control access to data and statistical techniques to mask data on individuals without destroying information about the group. A session on confidentiality at the 1983 Statistical Meetings focused on administrative rules, with the one exception of Spruill [3]. She proposed a statistical measure of the extent of disclosure and showed how it could be used to evaluate the effectiveness of masking techniques. Kincannon, a discussant in the session, predicted that in five years there would be another session on confidentiality focusing on statistical approaches to disclosure limitation rather than on administrative rules. Wendy Alvey deserves thanks for organizing this session ahead of schedule.

The papers presented in this session concern different aspects of statistical disclosure. Kim proposes a new method for limiting disclosure that promises to preserve statistical analyses better than current methods for limiting disclosure. Cox, Fagan, Greenberg and Hemmig study when and how cell suppression, controlled rounding and controlled perturbation can be implemented in two way tables. Strudler, Oh and Scheuren investigate empirically the effects of blurring, variable elimination and cell suppression on statistical analyses. Palley and Simonoff show that some common ways of limiting disclosure in databases can be defeated. Each of these topics — new ways to protect confidentiality, implementation of disclosure limiting techniques, effects of disclosure limitation on statistical analyses, and weakness of disclosure limiting techniques — is important and warrants further research.

The papers, however, illustrate a difficulty in statistical disclosure research. There is no common definition of disclosure, and not all researchers make their concept of disclosure explicit. Kim, who is concerned with microdata files, implicitly equates disclosure with unintentional release of the identity of the individual represented by a released data record. He mentions controlling the probability of disclosure, but he does not specify how the probability should be evaluated. Strudler, Oh and Scheuren also equate disclosure with unintentional release of the identity of a microdata record. Unlike Kim, they do not mention probabilities but they do apply measures of the extent of disclosure. Their measures are similar to those introduced by Spruill. Namely, for each "test" record in the masked file, a distance between the test record and each record in the source file is computed, and the percentage of test records that are closer to their corresponding source record than to any other source record is determined. This percentage multiplied by the fraction of source records that are released is a measure of disclosure. Although probabilities are not mentioned, the measures are related to probabilities.

Cox, Fagan, Greenberg and Hemmig consider tabular data rather than microdata, so the problem of associating an individual with a released record does not arise. Instead, the issue is controlling what can be inferred about an individual from released grouped data. These authors mention "the risk of disclosing confidential data", suggesting that there is both a probability and a loss associated with disclosure. The probability and loss models are not discussed, however.

In contrast, Palley and Simonoff allow information about aggregates as well as information about individuals to be confidential. For example, they suggest that a regression relationship between salary and employee characteristics, and not just the salary of a particular individual, may be confidential. Disclosure of confidential information about aggregates is measured in terms of how similar a model fit to released data is to a model fit to the source data. Disclosure of confidential information about individuals is measured in terms of how well a model estimated from released data predicts the source data. Palley and Simonoff's measures of disclosure, like those of Strudler, Oh and Scheuren, can be interpreted in terms of a probability model.

Plainly, different people have different intuitions about disclosure. These different intuitions lead to concern over completely different kinds of individuals. Kim, Cox et al. and Strudler et al. consider outliers to be more disclosure prone than non-outliers. The reason is that if a respondent is an outlier on variables available to the public, then the corresponding sensitive covariates of that respondent become available to the public when they are released with the public variables. In Palley and Simonoff's setup, however, data without outliers are more disclosure prone. The reason is that the regression of sensitive characteristics on non-sensitive characteristics can be inferred from the released data (threatening confidentiality of an aggregate), and the "typical" individual's sensitive characteristic can be predicted from the regression. The point is not that disclosure should be directed either towards or away from outliers. Rather, different interpretations of disclosure are possible and confusion is likely as long as intuition is not formalized.

I believe a framework that encompasses all the approaches to disclosure taken in this session can be constructed. The framework may be based on the work described in Duncan and Lambert [4] using probabilities and costs. Each paper in this session deals with probabilities, at least implicitly through summary statistics based on empirical distributions. Each also makes assumptions about the cost of disclosure. By including the cost explicitly, the controversy over whether outliers or typical observations are more important for disclosure may be analyzed in terms of costs. For example, if unusual characteristics are sensitive, then the cost of disclosing outliers may be so high that the risk of disclosure is unacceptable for outliers. Likewise, if the cost of revealing that a target individual is like everyone else is small, then the risk of disclosure may be acceptable for typical individuals even if their probability of disclosure is high. On the other hand, if releasing any information, no matter how unexceptional, about a target is undesirable, then the cost of disclosure is the same for all records and the risk of disclosure may be unacceptable for typical records.

Each paper in this session contributes to our understanding of disclosure; together they illustrate the wide variety of interpretations, problems and solutions. The challenge for the next session on confidentiality is to bring these different

insights into a unified framework so that they may reinforce each other.

NOTES AND REFERENCES

[1] Current address: AT&T Bell Laboratories, Rm. 2C-256, 600 Mountain Avenue, Murray Hill, N.J. 07974.

[2] Boruch, R. F. and Cecil, J. S. (1979). "Report from the United States: Emerging Data Protection and the Social Sciences' Need for Access to Data," in E. Mochmann and P. Muller, eds., Data Protection and Social Science Research. Springer-Verlag, N.Y., pp. 104-128.

[3] Spruill, N. (1983). "The Confidentiality and Analytic Usefulness of Masked Business Data." *1983 Proceedings of the Section on Survey Research Methods*, pp. 602-607, American Statistical Association.

[4] Duncan, G. T. and Lambert, D. (1986). "Disclosure-Limited Data Dissemination," Journal of the American Statistical Association, 81, 10-18 (in this volume).