STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Room Paper No. 1
English only

Topic I: Application of Statistical Disclosure Control Methodology and Software in Business Statistics and Social and Demographic Statistics

APPLICATION OF STATISTICAL DISCLOSURE CONTROL METHODOLOGY AND SOFTWARE IN BUSINESS STATISTICS AND SOCIAL AND DEMOGRAPHIC STATISTICS

List of key issues for discussion

Lawrence H. Cox
Associate Director, National Center for Health Statistics, USA

1. This topic is concerned with applications, but the papers offer much more. The two invited and eight contributed papers provide a stimulating array of applications, methodology, computational strategies, and empirical findings. Collectively, the papers focus attention in new and informed ways on six long standing issues in statistical disclosure limitation. We briefly summarize the papers within these issues.

2. *Identifying and measuring statistical disclosure*. Federal Statistical Office of Germany (2) deals with the important problems of primary disclosure risk assessment, the protection interval, and the problem of common respondents in tabular magnitude data (such as business data) subject to complementary cell suppression. Statistics Netherlands (3) discusses disclosure rules for tabular data and microdata and default values for disclosure rule parameters. Stricter rules are applied for public use microdata than for microdata for researchers. The Former Yugoslav Republic of Macedonia (7) presents experience with a legal framework for identifying statistical disclosure. National Agricultural Statistical Service (8) describes procedures based on a *p-percent disclosure rule* for agricultural data. EUSTAT (42) examines the consequences of selecting different values for the threshold parameter in a *threshold rule* for defining disclosure in frequency counts.

3. *Assessing and measuring disclosure risks.* Disclosure rules and disclosure risk are interrelated. In some situations, e.g., (3), (7), the disclosure rule is established and risk is based on the rule, e.g, likelihood of failing the rule. In other situations, e.g., (42), the rule may be set based on analytical or

empirical analysis of risk. Whereas the papers cited in the preceding paragraph focused mostly on the rule, the papers cited in this paragraph are concerned primarily with disclosure risk and its measurement. Institute for Employment Research (4) addresses measuring re-identification risk based on metrical and nearest-neighbor distances. Universitat Rovira i Virgili (5) examines risk from the standpoint of distance-based and probabilistic record linkage. Universidat de Valencia (9) develops a decision theoretic framework for controlling disclosure risk based on two factors: *discredit harm* to the national statistical office (NSO) based on intruder claims of achieving disclosure and actual *increase in intruder knowledge*. Carnegie Mellon University (10) provides a tool to assess disclosure risk as an algorithm for exact estimation of suppressed internal entries in certain multi-dimensional statistical tables.

4. *Methods for limiting statistical data disclosure*. Federal Statistical Office of Germany (2) presents a research plan for complementary cell suppression in large and complex tabulation structures that includes table-to-table protection and backtracking strategies. Universitat Rovira i Virgili (5) investigate the effectiveness of several methods for statistical disclosure limitation in microdata, several of which are new, e.g, *resampling* and *lossy compression*. Universitat Rovira i Virgili (6) examines theoretical properties of microaggregation. National Agricultural Statistical Service (8) reports using mathematical networks for complementary cell suppression in two-dimensional publication tables and related strategies for organizing table processing.

5. *Software for limiting statistical disclosure*. The research described in Federal Statistical Office of Germany (2) is related to the tau-Argus software. Statistics Netherlands (3) describes experience with the tau- and mu-Argus software. National Agricultural Statistical Service (8) describes experience with network-based software for cell suppression developed at the U.S. Census Bureau. EUSTAT (42) describes experience with tau-Argus.

6. *Computational issues and challenges*. Federal Statistical Office of Germany (2) is concerned with the many computational issues surrounding complementary cell suppression. Universitat Rovira i Virgili (6) addresses computational complexity issues for microaggregation. Results of this sort are extremely important in assessing the effectiveness and limitations of practical methods. Carnegie Mellon University (10) offers new computational methods for *disclosure audit* in a specific class of multi-dimensional tables.

7. *Assessing and limiting the effects of disclosure limitation on data analysis and usefulness*. This is an extremely important but difficult and under-explored area. Federal Statistical Office of Germany (2) discusses assigning *preferences* to candidate cells for complementary suppression. Statistics Netherlands (3) describes facilities for researcher access to original data *on-site* at the NSO. My organization, the National Center for Health Statistics, established one of the first Research Data Centers (RDC) for on-site access to confidential data. In addition, the NCHS RDC permits remote submission of SAS computer programs for analysis of original data, subject to review of computer programs and outputs. The NCHS RDC was discussed at the preceding meeting in Thessaloniki (Horm 1999). EUSTAT (42) deals with balancing disclosure protection and data utility through optimal choice of a threshold parameter for the case of frequency data. Universidat de Valencia (9)

provides an appealing theoretical framework for assessing interactions between disclosure risk and data utility.

8. The remainder of my remarks are comments and ideas for further research and general discussion stimulated by the papers. These are my own observations and are not intended to represent the policies or practices of the National Center for Health Statistics or any other organization. They are in no particular order.

9. The definition of statistical disclosure due to Dalenius states in essence that disclosure occurs if the release of a statistic S increases the intruder's knowledge about a respondent or reporting unit. The Dalenius definition implies that, except for the release of redundant or irrelevant information, disclosure almost surely will occur. This definition may appear impractical to some, but I think not. The Dalenius definition enables the concept that disclosure can be a matter of degree, e.g, divulging confidential information about two respondents is a greater disclosure than divulging the same information about only one of the respondents, or that divulging a narrow estimate of a confidential quantitative attribute is greater disclosure than divulging a broad estimate. Upon this quantitative framework, one can build a model for disclosure risk incorporating both *the degree of disclosure* and the *likelihood of disclosure*. Universidat de Valencia (9) presents related work.

10. *Re-identification studies*, as presented in Institute for Employment Research (4) are extremely important for measuring and assessing disclosure risk that should be conducted on a routine basis by NSOs, but sadly are not. Subsequent to the Thessaloniki meeting, John Horm prepared a preliminary unpublished study dealing with evaluation of several population thresholds against re-identification in microdata from health surveys.

11. It would be interesting to see the graph theoretic methods presented in Carnegie Mellon University (10) applied to the complementary cell suppression problem implied by their results: when exact bounds are too narrow, how should marginal totals be selected for supopression to ensure acceptable disclosure risk?

12. Institute for Employment Research (4) and Universitat Rovira i Virgili (5) are concerned with record linkage. It would be interesting to see an examination of the methodological and computational requirements of record linkage both for practical purposes and theoretically.

13. Federal Statistical Office of Germany (2) addresses the *problem of common respondents*, also known as the *multi-cell* disclosure problem. Standard models for complementary suppression limit disclosure only at the *cell level*, viz., are based on providing at least the minimum numeric protection necessary to ensure that the union of the disclosure cell with a non-disclosure cell containing <u>different respondents</u> will be a non-disclosure cell. These methods ensure necessary but not sufficient conditions that, e.g, respondent data cannot be estimated closer than p-percent. More complicated mathematical models, illustrated in Cox (1999), and improved computational strategies are needed to address the common respondent problem fully.

14. Carnegie Mellon University (10) provides simple formulae for obtaining exact estimates of suppressed internal entries in multi-dimensional tables, subject to the condition that the released marginal totals constitute a set of sufficient statistics for a *decomposable graphical model*. I have been investigating the class of log-linear models representable as mathematical networks, resulting in the same formulae for exact bounds (Cox 2000b). I am exploring further relationships between these two classes, and extension to the case of reducible models. In particular, decomposable graphical models are network, as are the k-dimensional dichotomous tables on which the algorithm of Section 4 of Carnegie Mellon University (10) is based. Moreover, this work reveals an interesting application of network models to Markov Chain Monte Carlo computation for decomposable log-linear models.

15. As above, bounding problems can be approached from several vantage points in mathematical science. Iterative proportional fitting (IPF) is central to the existence of feasible multi-dimensional tables with given marginals. Preliminary unreported results of a colleague indicate that, for feasible tables, iterative proportional fitting tends to provide close estimates of original suppressed internal entries. IPF results in maximum likelihood estimates, but still it is not clear if (and why) entries removed by mathematical algorithms should be nearly reconstructible by a statistical algorithm. This is important to investigate. So, too, is the use of Bayesian methods based on prior distributions for suppressed entries to develop probability distributions on feasible values for suppressed internal entries.

References

Cox, L.H. (1999). "Invited paper: some remarks on research directions in statistical data protection." *Proceedings of Statistical Data Protection '98, Lisbon*. Luxembourg: EUROSTAT, 163-176.

_____ (2000a). "On properties of multi-dimensional statistical tables." Submitted.

_____ (2000b). "Representing log-linear models as mathematical networks." Lecture notes.

Horm, J. (1999). "National Center for Health Statistics approaches to protection and release of microdata." **Statistical Data Confidentiality**. Luxembourg: EUROSTAT, 75-83.