

# University of Maryland Statistics Consortium Seminar

**Speaker:** Dr. Lawrence H. Cox, Associate Director, National Center for Health Statistics

**Title:** Resolving confidentiality and data quality issues for tabular data

**Location:** 2205 Lefrak Hall, University of Maryland, College Park

**Date and Time:** 3:00-4:00PM, March 23, 2005, Wednesday

A reception will follow the talk. Please visit the following website for direction, parking and updated information:

*<http://www.statconsortium.umd.edu>*

*Abstract:* Traditional methods for statistical disclosure limitation in tabular data are cell suppression, data rounding and data perturbation. Each method is applicable to count data whereas only cell suppression has been effective for magnitude data. Cell suppression can create obstacles for the data user because it eliminates otherwise useful data, thwarting straightforward and complete data analysis. Cell suppression can frustrate the statistician because the suppression mechanism is unrelated to standard probability models, e.g., missing-at-random, with the result that suppressed tables are not amenable to familiar imputation methods. The deteriorating effects of cell suppression on data quality and completeness are largely unmeasured. Consequently, data quality characteristics of suppressed tables are poor. Cell suppression is equally a problem for the data protector because suppression is a theoretically and computationally difficult problem (NP-hard). In short, cell suppression is nobody's favorite methodology but for decades provided the only viable avenue for confidentiality protection in tabular magnitude data.

Recently, an alternative to cell suppression emerged, *controlled tabular adjustment* (CTA). CTA replaces tabular cells that represent disclosure by safe values, uses mathematical programming to adjust remaining values to rebalance tabular equations, and (nearly) optimizes any of a range of global and local measures of data "closeness". The latter step addresses data quality, but only in part and, until recently, not in a statistical way. Current research discussed in this talk is aimed at preserving distributional properties of original data in released (adjusted) data. For univariate data, we seek to ensure that means and variances of adjusted data are reasonable approximations of means and variances of original data and that original and adjusted data exhibit high positive correlation. For multivariate data, in addition to univariate properties, we seek to ensure that covariances, correlations and regressions between two variables exhibited by adjusted data approximate those between original data. We do so using mathematical programming. These problems can be formulated exactly in the context of nonlinear programming, but then are unlikely to be computable. Instead, we present approximations to exact formulations based entirely on linear programming, which consequently are easy to understand, implement and enhance. This research is one example of the possibilities for applications-oriented mathematical sciences research at the statistics/operations research interface.