Dissertation Defense

Mandi Yu

Friday, July 25

3:00pm

368 ISR

and

1208 LeFrak

DISCLOSURE RISK ASSESSMENTS AND CONTROL

Recent advances in technology dramatically increase the volume of data that statistical agencies can gather and disseminate. The improved accessibility translates into a higher risk of identifying individuals from public microdata, and therefore increases the importance of the evaluation of disclosure risk and confidentiality control. This dissertation addresses three related but distinct research questions in statistical data confidentiality.

The first study concerns the evaluation of disclosure risk for microdata when an intruder attempts to identify survey respondents by linking data records with a large external commercial data file based on a set of common variables. The dependence of disclosure risk to the commercial data coverage, the accuracy of the common identification information, and the amount of identification information to which an intruder accesses, is discussed theoretically and empirically tested using an experiment.

The second study presents a practical implementation of fully-imputed synthetic data approach for a large, complex longitudinal survey as means of protecting confidentiality, following the initial proposal by Rubin (1993) and Little (1993). The imputation uses separate semiparametric algorithms for continuous, binary and categorical variables. A new combining rule of synthetic data inference is proposed to account for the uncertainty due to simultaneously imputing item-missing data and generating synthetic data. The loss of data utility is evaluated via the use of a propensity score approach in addition to three information loss metrics.

The third study extends this fully-synthetic data approach to cope with situations where small area statistics are essential important. This research is the first in the statistical disclosure control literature to consider small area statistics. The goal is to create synthetic data with enough geographical details to permit small area analyses, which otherwise is impossible because such geographical identifiers are usually suppressed due to disclosure control. A Bayesian framework for appropriate small area models is proposed to generate synthetic microdata from the predictive posterior distributions. Two simulation studies and one empirical illustration are used to evaluate this approach.