STATISTICS SEMINAR

Samuel S. Wu, Ph.D.
University of Florida
Thursday, November 6, 2014
Dickens Hall, Room 207, 4:00-5:00 pm
Refreshments: Dickens 108, 3: pm

New Data Collection Methods with Privacy Protection

A major obstacle that hinders medical and social research is the lack of reliable data due to people's reluctance to reveal confidential information to strangers. Fortunately, statistical inference always targets a well-defined population rather than a particular individual subject and, in many current applications, data can be collected using a web-based system or other mobile devices. These two characteristics enable us to develop new data collection methods with strong privacy protection. These new technologies hold the promise of removing trust obstacle, promoting objective data collection, allowing rapid data dissemination, and helping unrestricted sharing of big data.

The new methods ensure that the raw data stay with research participants and only masked data are collected, which can be distributed and shared freely. The new method, called triple matrix-masking, offers strong privacy protection with an immediate matrix transformation at time of data collection so that even the researchers cannot see the raw data, and then further uses matrix transformations to guarantee that the masked data will still be analyzable by standard statistical methods. A critical feature of the method is that the keys to generate the masking matrices are held separately, which ensures that nobody sees the actual data. Also, because of the specially designed transformations, statistical inference on parameters of interest can be conducted with the same results as if the original data were used, hence the new method hides sensitive data with no efficiency loss for statistical inference of binary and normal data, which improves over Warner's randomized response technique. In addition, we will present some variations of the method and their properties regarding data quality assurance and data security.

Keywords: Orthogonal transformation, Privacy-preserving data collection, General linear model, Contingency table analysis, Logistic regression.