

STATISTICS COLLOQUIUM

THURSDAY, APRIL 26, 2012

TALK: 10:00 AM — SCIENCE CENTER RM. 300H

“A Discussion of Uncongeniality for Synthetic Data”

**Bronwyn Loong
Statistics Department
Harvard University**

ABSTRACT

In the multiple imputation literature, uncongeniality refers to differences in types of input between the imputer, typically an agency, and the analyst. When releasing synthetic data in place of observed values to protect data confidentiality, the imputer and external user are always separate bodies, hence the potential for uncongeniality always exists.

In this talk, we propose a definition of congeniality for multiple imputation for synthetic data. Our definition is motivated by common examples of uncongeniality, specifically ignorance of the original survey design in analysis of fully synthetic data, and situations when the imputation model and analysis procedure condition upon different sets of records. We illustrate the definitions with empirical results from synthesis of a large-scale health survey. We conclude that our definition assists the imputer to identify the source of a discrepancy (or potential discrepancy) between observed data set and synthetic data analytic results. This is important to gain public confidence that valid statistical inference can be obtained given synthetic data. Even so, it is impossible for imputers to satisfy congeniality for all potential future analyses and therefore, it is also important for analysts to be educated on uncongeniality and its statistical implications. Based on our empirical work, our recommendation is for analysts to use synthetic data primarily for exploratory data analysis purposes.