Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

Multiple imputation methods for unknown stage at diagnosis in cancer data

Pradeep Singh (Southeast Missouri State University)

Abstract: The National Cancer Institute and most states keep a cancer data registry so that it can be used by researchers and policy makers to make better healthcare decisions. This data can have missing observations for one or more variables. In particular, the correct stage at diagnosis is sometimes missing from the data due to various reasons. To use the data, different strategies have been used. Researchers often delete individuals from the study who had missing values from even one variable. Another method is to impute the missing values. There are several methods proposed to impute missing values of quantitative variables. But for categorical variables, there have been few methods proposed. Van der Palm, et al. [2016], compared four imputation methods for categorical data. Zhou et al. [2017] has proposed a nonparametric multiple imputation method using the nearest-neighbor approach. This study applied the nonparametric multiple imputation method proposed by Zhou et al. [2017] and a parametric multiple imputation method to lung adenocarcinoma data from the National Cancer Institute. Lung adenocarcinoma is a type of non-small cell lung cancer that typically forms on the outside of the lungs. A Monte Carlo study was done to compare these methods with respect to imputation bias. The study also compared the effect of different levels (10%, 20%, 40%) of missingness, different sizes of the sample, and different fits of the model on these multiple imputation methods.

Wednesday, March 17 at 4:00 PM in Zoom