Yale SCHOOL OF PUBLIC HEALTH

Biostatistics

Generating Poisson-Distributed Differentially Private Synthetic Data

Harrison Quick, PhD

Assistant Professor

Department of Epidemiology and Biostatistics

Drexel University

12:00 Noon Eastern time, Tuesday, October 6, 2020

Virtual seminar via Zoom

ABSTRACT

The dissemination of synthetic data can be an effective means of making information from sensitive data publicly available with a reduced risk of disclosure. While mechanisms exist for synthesizing data that satisfy formal privacy guarantees, these mechanisms do not typically resemble the models an end-user might use to analyze the data. More recently, the use of methods from the disease mapping literature has been proposed to generate spatially referenced synthetic data with high utility but without formal privacy guarantees. The objective for this paper is to help bridge the gap between the disease mapping and the differential privacy literatures. In particular, we generalize an approach for generating differentially private synthetic data currently used by the U.S. Census Bureau to the case of Poisson-distributed count data in a way that accommodates heterogeneity in population sizes and allows for the infusion of prior information regarding the underlying event rates. Following a pair of small simulation studies, we illustrate the utility of the synthetic data produced by this approach using publicly available, county-level heart disease-related death counts. This study demonstrates the benefits of the proposed approach's flexibility with respect to heterogeneity in population sizes and event rates while motivating further research to improve its utility.