

Catalogue no. 11-522-XIE

**Statistics Canada International  
Symposium Series - Proceedings**

**Symposium 2006 :  
Methodological Issues in  
Measuring Population Health**

2006



**Statistics  
Canada**

**Statistique  
Canada**

**Canada**

## **Evaluation of Methods for Outlier Detection and Treatment in the U.S. Survey of Occupational Illnesses and Injuries**

John L. Eltinge<sup>1</sup>

### **Abstract**

The U.S. Survey of Occupational Illnesses and Injuries (SOII) is a large-scale establishment survey conducted by the Bureau of Labor Statistics to measure incidence rates and impact of occupational illnesses and injuries within specified industries at the national and state levels. This survey currently uses relatively simple procedures for detection and treatment of outliers. The outlier-detection methods center on comparison of reported establishment-level incidence rates to the corresponding distribution of reports within specified cells defined by the intersection of state and industry classifications. The treatment methods involve replacement of standard probability weights with a weight set equal to one, followed by a benchmark adjustment.

One could use more complex methods for detection and treatment of outliers for the SOII, e.g., detection methods that use influence functions, probability weights and multivariate observations; or treatment methods based on Winsorization or M-estimation. Evaluation of the practical benefits of these more complex methods requires one to consider three important factors. First, severe outliers are relatively rare, but when they occur, they may have a severe impact on SOII estimators in cells defined by the intersection of states and industries. Consequently, practical evaluation of the impact of outlier methods focuses primarily on the tails of the distributions of estimators, rather than standard aggregate performance measures like variance or mean squared error. Second, the analytic and data-based evaluations focus on the incremental improvement obtained through use of the more complex methods, relative to the performance of the simple methods currently in place. Third, development of the abovementioned tools requires somewhat nonstandard asymptotics that reflect trade-offs in effects associated with, respectively, increasing sample sizes; increasing numbers of publication cells; and changing tails of underlying distributions of observations.

---

<sup>1</sup>Bureau of Labour Statistics, Washington, DC, USA