Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

PROTECTING CONFIDENTIALITY WHILE PRESERVING QUALITY OF PUBLIC USE MICRO DATA

Avinash C. Singh, Moshe Feder, George Dunteman, and Feng Yu[1]

ABSTRACT

The analogy between sampling from a finite population to reduce enumeration cost, and that from a data-base to reduce disclosure risk, is exploited to propose a general method of disclosure treatment of public-use micro data while maintaining analytical utility. Thus the steps of optimal sample design, edit and imputation, unit nonresponse adjustment, and finally poststratification used in survey sampling can be modified to suit data-base sampling consisting of four steps termed micro agglomeration, subsampling, substitution, and calibration. Clearly there is no need of edit, imputation and unit nonresponse adjustment because all the information is available from the full data-base. However, there is a need of another form of imputation in which a random sample of unique records from the data-base is treated as missing, and their values are substituted from similar records. This sampling for substitution introduces uncertainty about the uniqueness of a record, similar to sampling from the data-base that introduces uncertainty about the known presence of a record, and can be done optimally to minimize disclosure risk while controlling MSE of key estimates. Empirical results based on experience with the National Household Survey on Drug Abuse will be presented.

Paper not received

[1] Research Triangle Institute, USA