Introduction to Controlled Data Swapping

The IES confidentiality standards generally require projects to introduce an additional measure of uncertainty into the data using controlled random swapping. The *DataSwap* software was developed and enhanced in an effort to ease the implementation of this procedure. *DataSwap* provides NCES with standardized approaches to perturb data, and evaluate the impact on data utility. Through *DataSwap*, controlled random swapping is used in response to re-identification threats. Krenzke et al (2006) provide ways to approach the use of data swapping as applied to cross-sectional surveys, survey data with hierarchical structure, and longitudinal data structures. To balance the need to address re-identification risk while retaining data utility, the *DataSwap* software allows the user to evaluate the impact of data swapping on data quality. Kaufman, Seastrom and Roey (2005) discuss the impact of controlled random swapping on data quality using data from the Trends in Mathematics and Sciences Survey. Dohrmann et al. (2009) provide the measures of global utility that is embedded in the *DataSwap* software.

Benefits

The basic idea of controlled random swapping is to protect a database by interchanging, or "swapping" values of one or more variables between records. A benefit of swapping and using any other random perturbation method, is that there is added uncertainty by creating false positive matches that are undistinguishable from true positive matches by the data intruder. The intruder can match up the data with the known key characteristics or with an external file, but never knows if the matched data are true. Note that this type of uncertainty cannot be measured. The *DataSwap* approach to data swapping is a random perturbation method, and therefore it inherits this type of uncertainty for which the resulting risk level cannot be measured. Another benefit of using swapping as a statistical perturbation technique is that it maintains the unweighted univariate distribution of each variable while still introducing uncertainty about the identity of records. A data intruder does not know which variables or records contain swapped information, so randomized swapping helps, to make re-identification of a record uncertain for the intruder.

Other approaches

There are two broad classes of statistical confidentiality treatments to microdata: synthetic and perturbation approaches. Synthetic data approaches involve producing fully synthetic datasets (Rubin, 1993) or partially synthetic datasets that are mixtures of actual and multiply-imputed values (e.g., Liu and Little, 2012). Synthetic data approaches typically replace original values with draws from appropriate probability distributions in a way that aims to retain the essential statistical features of the original data, including multivariate associations. Most synthetic approaches rely on the multivariate relationship between the target variables (variables to be masked) and other variables in determining the synthetic values. When missing data patterns are non-monotone (e.g., some may refer to as "Swiss cheese") as typical in surveys, to maintain relationships between variables, a sequential approach (Raghunathan, 2001) is used where synthetic values are drawn for Variable 1 from the posterior predictive distribution of observed data on predictor variables. Then synthetic values are drawn for Variable 2 from the posterior predictive distribution that includes the synthetic values from Variable 1 and observed values from other predictor variables. The process continues until all variables targeted for masking are synthesized. The process is circular and ends based on convergence rules or when a predetermined number of cycles are conducted.

Perturbation approaches involve applying a controlled random treatment procedure (e.g., swapping) to replace a subset of the original data values by other values, with the aim of introducing just enough noise or uncertainty into the microdata to reduce the disclosure risk to an acceptable level. Perturbation methods are sometimes referred to as blank-and-impute and can control change from the original values. Several perturbation approaches use the sequential approach outlined above to maintain multivariate associations. The approach by Krenzke, Li and McKenna (2017) uses a perturbation procedure that is based on a sequential imputation procedure to replace data values in the American Community Survey data that were used to generate special tabulations called the Census Transportation Planning Products.

For the sequential regression or perturbation approaches, a key feature is to select variables for the treatment model from a large pool of variables. This ensures that the most important variables are used in the treatment process, and in doing so, multivariate associations have a better chance of being retained. When applying controlled random swapping, one needs to keep the concept of retaining multivariate associations at the forefront of the planning stage. Prior to implementation, a search for variables to be included in the swapping process is critical. A carefully designed process is needed and the setup within *DataSwap* can make a difference in the impact on data utility. Therefore, after the approach is described, and the process presented, some helpful tips are provided to help guide the process toward a reduced impact on data utility.

Much has been said about transparency of the process to the data user. This is an important topic. NCES' policy is to not provide the swapping variables or the swapping rate. However, the algorithm is publicly available. Those doing the swapping must ensure the process will not impact the conclusions made from the data.

References

Dohrmann, S., Krenzke, T. Roey, S., and Russell, N. (2009). Evaluating the impact of data swapping using global utility measures. Proceedings of the Federal Committee on Statistical Methodology Research Conference.

Kaufman, S., Seastrom, M., and Roey, S. (2005). Do disclosure controls to protect confidentiality degrade the quality of the data? Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Government Statistics. Alexandria, VA: American Statistical Association.

Krenzke, T., Li, J., and Mckenna, L. (2017). Producing multiple tables for small areas with confidentiality protection. *Journal of the International Association of Official Statistics*, 33(2), 469-485. doi: 10.3233/SJI-160259

Krenzke, T., Roey, S., Dohrmann, S., Mohadjer, L., Haung, W., Kaufman, S., and Seastrom, M. (2006). Tactics for reducing the risk of disclosure using the NCES DataSwap software.

Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.

Liu, F. and Little, R. (2012). Multiple imputation and statistical disclosure control in microdata. Joint Statistical Meetings Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association; 2012: 2133–2138.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology*. 27:85–96.

Rubin, D. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*. 9:462–468.