Quantifying re-identification risk

The variables common to an adversary's knowledge (e.g., factual data items such as demographics) or common to external files, can be combined in an attempt to identify individuals or establishments. Only a few variables are necessary to be used in combination to result in a sample unique.

Consider the expected number of sample uniques being also population uniques expressed as:

$$GlobalRisk = \sum_{SU} P\left(F_k = 1 \mid f_k = 1\right)$$

where SU is the set of sample uniques, f_k is the sample frequency in cell k, and F_k is the population frequency in cell k. Hundepool et al. (2012) discuss disclosure in a microdata context as a re-identification operation that is achieved by an intruder when comparing a target individual or entity in a sample with an available list of units (external file) that contains direct identifiers (e.g., name and address), plus a set of indirect identifying variables. Re-identification occurs when the unit in the released file and a unit in the external file belong to the same individual in the population. The risk also exists to a certain extent since a 'nosy neighbor' may know a handful of facts about a person or entity and could search the file to find the person or entity. In practice, F_k needs to be estimated. Investigation into the use of models to provide more stable estimates of risk has been conducted by researchers. Skinner and Shlomo (2008) provide an improved risk measure using log-linear models. The assessment of re-identification risk based on the log-linear modeling assumes that the intruders attempt to link the sampled cases, especially the sample uniques, to those in the population using a set of known characteristics. To do so, we assume approximately ten variables are known by the intruders accurately. The characteristics can be publicly available or obtainable in external sources, such as geographical variables, demographic variables (e.g., age, sex, race/ethnicity), and sensitive attributes (e.g., disability, income). If there are design variables that lead to large variations in selection probabilities, they should be included in the log-linear models as well. For the purpose of the risk assessment, some categories of the variables can be combined assuming it is not likely for data intruders to know the original detailed categories. Different assumptions on the known information by the intruders may result in different estimates of risk. The log-linear models are helpful to quantify the re-identification risk of the file to be released.

When setting up a risk assessment, the average cell size can be computed as the sample size divided by the number of cells formed by crossing the indirect identifying variables (the cells with zero sampled cases are accounted for in the denominator). The more indirect identifying variables used in the risk assessment and/or the more categories that the indirect identifying variables have, the smaller the average cell size. In general, the estimated risk increases as the average cell sizes decrease, with other conditions being the same. In other words, the more accurate and detailed information that the intruder may know about the respondents, the higher the re-identification risk.

The risk measure results using the loglinear model are estimates of the re-identification rate among the sample records, i.e., the percentage of the sample uniques that are estimated to be population uniques, with the possible range from 0 percent to 100 percent, where 0 percent indicates no risk of disclosure and 100 percent indicates that each record is unique in the population. In general, the risk can be estimated using log-linear model with all 2-way interactions, using survey weights. The cell average sample fraction should be used to calculate the risk estimates and goodness of fit criteria when stratifiers are included in model covariates (Skinner and Shlomo, 2008; Li, Li, Krenzke, submitted).

The determination of high or low risk is an internal rule-of-thumb, and NCES may consider or determine other risk thresholds for their data products. The risk value is based on a particular metric, and it depends on how identifiable the variables are among those used, the number of variables, and the number of categories used. A heuristic rule-of-thumb loosely follows 1% or less as very low risk, 3% or less as low, 6% or less as moderate, and over 6% as high, however, it is dependent on the level of access, identifiability of the variables, and the harm that a breach may cause.

Accessing the Re-identification Risk Quantification Software

Software for log-linear models for disclosure risk: The R package sdcnway is available on the CRAN network at https://CRAN.R-project.org/package=SDCNway.

References

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., and de Wolf, P.-P. (2012). *Statistical Disclosure Control.* Chichester, UK: John Wiley & Sons. doi: 10.1002/9781118348239

Li, L., Li, J., and Krenzke, T. (submitted). Reliance on goodness-of-fit criteria toward measuring re-identification risk using log-linear models.

Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of American Statistical Association*, 103, 989–1001.