A Post-Aggregation Error Record Extraction Based on Naive Bayes for Statistical Survey Enumeration

Kiyomi Shirakawa, National Statistics Center, Tokyo, JAPAN
e-mail: kshirakawa@nstac.go.jp

Abstract

At the pre-aggregation stage of micro-editing in the National Statistics Center (NSC), an optimal editing of individual data has been implemented by consistency check and range check for each survey. In particular, some errors in numeric data items are logically detected based on a probability distribution of fixed parameters of the sample mean and standard deviation. However, when an acceptance region (range) for the error detection is narrow, a lot of correct data are detected. On the other hand, when the range is wide, many errors are not detected. In general, the problem associated with the range is revealed during a data review process at the post-aggregation stage. In the past, it was customary to edit manually. In order to solve this problem, it is necessary to extract the errors in the cell of the statistical table after aggregation. Therefore, we propose the introduction and systematization of naive Bayes that can treat a parameter as a random variable. In this method, the errors can be filtered by subjective probability. Thus, it is not a strict range check compared with objective probability, such as Smirnov-Grubbs' test. The filtering in subjective probability is dependent on a combination of several items and prior probability by unit of records. In addition, the validation test of the records may also include individual data with missing values. In this paper, we assess the error extraction method that focuses on sales variable in aggregated cells of the statistical table. As a result, it is possible to extract the errors in the point of view that is different from micro editing, and that the data editing process is automated.

Keywords: cell data, cross-tabulation table, filtering, inliers, subjective probability