A Method for Detecting Outliers in Survey Data by Ratio Tests

Cheng Bangwen*, Yang Hongjin, Shi Linfen, Wang Yali, Xu Ji
School of Management, Huazhong University of Science and Technology,
Wuhan, China Chengbw@hust.edu.cn

Detecting abnormal data in original survey data is an important and difficult work in statistical surveys. In socio-economic statistics, the survey items most frequently encountered are "scale indicators" reflecting the size or scale of research objects, such as yield, output value, fund, personnel, and assets. For this kind of indicators, although outliers in original survey data from samples can be detected by univariate methods such as distribution tests or boxplots, the method's power of detecting outliers is limited due to masking and swamping effects during the process of searching for outliers. The main reason for this situation is that the method does not exclude the influence of the scale of research objects on data. In order to overcome this defect, a new method based on ratio tests is proposed. This paper discusses the basic principles and basis of the method, explores the mechanism, condition, and the way to improve detection effectiveness. An empirical study based on the original data of government research institute survey of China has been taken. Starting from the essential attribute or characteristic of research objects, the method of ratio tests uses distribution tests or boxplots of ratio data to detect outliers, which can eliminate the influence of individual scale on outlier detection with the help of the ratio of the two scale indicators. In the paper, it has been shown that when a certain condition is satisfied the ratio indicator will have lower coefficient of variation so that outliers can more easily be detected, and when the conditions are not satisfied the ratio test result will become worse. Reasonably selecting ratio indicators is the key to effectively detect outliers. The proposed method is also suitable for data with probability distribution unknown. Empirical study shows that the method is simple, practical, easy to use, and has application value.

Key Words: Quality of statistical data, probability distribution test, coefficient of variation, lognormal distribution, boxplot