AUTOMATED DETECTION OF OUTLIERS

Jelena Voronova
Central Statistical Bureau/University of Latvia, Latvia, Jelena.Voronova@csb.gov.lv

Anomalies, or outliers, can be a serious issue applying statistical techniques. Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, or none. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

When applying standard sampling weights to outliers, estimates are believed to be distorted. It is important to recognise that the distortion is, in principle, the result of a large sampling error and not a bias. Isolating outliers may also have a positive impact on the results of data analysis. Simple statistical estimates, like sample mean and standard deviation can be significantly biased by individual outliers that are far away from the middle of the distribution.

Since outliers usually have a huge impact on estimates, outlier detection and their treatment are important elements of statistical analysis. This is true especially when estimation is carried out at a low level of aggregation. In the case of small sample sizes, outliers can affect variance. Even if the sample size is large, the influence of an outlier can significantly increase the variance resulting in a decreased efficiency of estimation.

Outliers can be representative (representing other population units similar in value to the observed outliers) or non-representative (unique in the population). Representative outliers should be handled in the survey estimation process, by the use of outlier resistant or robust estimation procedures.

There are methods of dealing with outliers in a finite population, apart from removing them from the dataset. In many business surveys it is a relatively common practice, reducing the weights of outliers (trimming weight), by setting the survey weight equal to one.

In business surveys, the distribution of variables is often highly skewed, resulting in sample observations that differ substantially from the majority of observations in the sample. Dealing with business Short Term Statistics (STS) we cannot call in questions economic indicators, giving to NSI by the enterprises. Procedures are developed for detecting outliers in continuous univariate data. Talking in account results of the analysis of the variance and computed estimates in the time series, sometimes we could identify not detected anomalies in the dataset. Having the limitations in a sampling error and using the automated process of detecting and isolating outliers, we are dealing with problem of automated outlier detection in STS datasets with the small sample size and high level of aggregation.

References

Chambers, R. L. (1986), Outlier Robust Finite Population Estimation. Journal of the American Statistical Association.

Cox, B.G., D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, P.S. Kott, eds. 1995. Business Survey Methods. John Wiley and Sons, New York.

Eltinge, J. and Cantwell, P. (2006), Outliers and Influential Observations in Establishment Surveys. Paper prepared for presentation to the Federal Economic Statistics Advisory Committee (FESAC).

Ren, R. and Chambers, R. L. (2002a), Outlier Robust Methods: Outlier Robust Estimation and Outlier Robust Imputation By Reverse Calibration. Report for Euredit.