OUTLIER DETECTION METHODS FOR BUISENESS SURVEYS

Anton Tovchenko
State Statistics Service of Ukraine, Ukraine, A.Tovchenko@ukrstat.gov.ua

Olexiy Tkachenko
State Statistics Service of Ukraine, Ukraine, tom@ukrstat.gov.ua

The quality of sample enterprise survey, unlike sample social survey, greatly depends upon the number of outliers and the methods of outlier detection (that is, enterprises with big values of relevant indicator). Enterprise statistics has a number of traits:

1. The distribution of observed value is not normal, it is hyperbolic, and so the values of about 5% of enterprises can make up to 90% of total value. In addition, it makes using parametric methods unfeasible (e.g. Pirson correlation or dispersion analysis).
2. In enterprise statistics the stratified sampling is usually used (stratum=domain). The accuracy of estimation of domains is of critical importance. In addition, the sum of values across different domains may greatly vary (10 times or more).
3. While building sample design, the real values are usually known across all the general population, because the census usually precedes the sample survey.

Due to mentioned traits, there is a question as to how many outliers to detect and with which method. In addition, the not only the accuracy of estimation of total population must be assured, but the estimation of all the relevant domains as well.

The conclusion on the quality of the sample was based on the value of dispersion of estimation of total and coefficient of variation of estimation, total and across domains. The quantity of enterprises in the sample of about 20% of the total population was deemed necessary. The testing of the design consisted of the following steps:

1. Outlier detection individually for each domain.
2. Allocation of the rest of the enterprises (total quantity minus the quantity of outliers) across domains using Neumann allocation.
3. Sampling.
4. Calculation of coefficient of variation of estimation (total and for each domain).

State Statistics Service of Ukraine specialists tested the following methods of outlier detection:

1. Iterative parametric method: "n-sigma" coefficient for standard deviation was changed from 1 to 5 by step 0,2.
   (parameter > mean + coefficient*(standard deviation))
2. Iterative non-parametric method: coefficient for interquartile range was change from 1 to 5 by step 0,2.
   (parameter > median + coefficient*(interquartile range))
3. Iterative minimization of coefficient of variation in each domain to a value that was changed form 300% to 100% by 10%.
   (sort by relevant indicator in descending order, iteration: "calculate CV, if CV is greater than specified value - top enterprise is an outlier", continue until CV becomes lesser or equal to specified value)

Preliminary results:

1. Using parametric method leads to the worst quality (especially in some domains), what is more, the optimal coefficient for standard deviation varies in range from 1 to 2, which is contrary to "heuristic" three.

2. Using non-parametric methods gives better results across domains; the optimal coefficient for interquartile range is in range from 4 to 5.
3. Using Iterative minimization of coefficient of variation in each domain gives the best results.

The presentation and the paper will illustrate and describe the results with the conclusion of which method was better to detect outlier.

References

Kish L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley & Sons.

Hansen M.H., Hurvitz W.N. & Madow W.G. (1953). *Sample survey methods and theory*. New York: John Wiley & Sons.