

Treatment of Outliers

In the *Tankan* (Short-Term Economic Survey of Enterprises in Japan), population estimation -- a statistical method -- is applied to aggregation of quantitative data such as sales, by strata. To further improve statistical accuracy, the *Tankan* has introduced the treatment of outliers since the December 2010 survey, under which outliers reported from sample enterprises are treated in a special way for calculating estimates for population total. Details on the procedures, including the detection and the treatment of outliers, are described below.¹

1. Method of Outlier Detection

(1) y_{it} , which denotes the degree of influence of values reported from each enterprise on year-on-year rates of change or revision rates, is calculated for the six categories -- two industries (manufacturing and nonmanufacturing) times three capital sizes (large, medium-sized, and small enterprises) -- as follows:

$$y_{it} = \frac{\text{Degree of change in weighted data}}{\left| \begin{array}{l} \text{Estimate for population total of one of the six categories} \\ \text{in the previous fiscal year (or previous survey)} \end{array} \right|} \times 100$$

$$= \frac{w'_i(x_{it} - x_{it-1})}{|M_{jt-1}|} \times 100$$

x_{it} : Value reported from enterprise i at time t

x_{it-1} : Value reported from enterprise i at time $t-1$

w'_i : Adjusted weight of x_i *

M_{jt-1} : Estimate for population total of category j ($j=1, \dots, 6$) at time $t-1$

* To accurately measure the degree of change in weighted data, despite the difference in the number of sample enterprises between time t and $t-1$, w'_i is computed based on the same number of sample enterprises. For this purpose, when some enterprises provide no responses in either of two surveys (at different points in time), missing values are substituted by the average values of each stratum for population estimation to which they belong.

¹ For details of outliers, see "Treatment of Outliers in Business Surveys: The Case of Short-term Economic Survey of Enterprises in Japan (*Tankan*)" (September, 2010).

(2) Next, z_{it} , the indicator of the degree of divergence based on y_{it} , is calculated as follows:

$$z_{it} = \begin{cases} \frac{y_{it} - d_j^{99}}{D_j} & \text{if } d_j^{99} \leq y_{it} \\ \frac{d_j^1 - y_{it}}{D_j} & \text{if } y_{it} \leq d_j^1 \\ 0 & \text{if } d_j^1 < y_{it} < d_j^{99} \end{cases}$$

If this indicator, z_{it} , exceeds the threshold value ($C=50$) -- which is set in advance based on empirical analyses using past data -- the reported value is detected as an outlier.

- d_j^1 : 1 percentile² of y_{it} in category j
- d_j^{99} : 99 percentile of y_{it} in category j
- D_j : Distance between d_j^{99} and d_j^1 ($d_j^{99} - d_j^1$)

The treatment of outliers is applied to the following five items: "Sales," "Current Profits," "Net Income," "Fixed Investment," and "Software Investment." Figures for "Exports" and "Land Purchasing Expenses" -- which are part of "Sales" and "Fixed Investment," respectively -- are treated as outliers only when those for "Sales" and "Fixed Investment" are detected as outliers.

2. Method of Outlier Treatment

Detected outliers are treated as missing values, since they do not represent the population; using such data for aggregation is deemed inappropriate. In this case, with the rule of the missing value imputation applied, these values are substituted by the most recent figures answered by enterprises ($x_{it} = x_{it-1}$), excluding the influence of the outliers. Once figures are regarded as an outlier, this treatment is continued until sample enterprises of the *Tankan* are revised.

² The 1 percentile is defined as the value below which 1 percent of observations can be found.