# Processing of Influential Values in Surveys

Cyril Favre-Martinoz & Thomas Deroyon

***Résumé* — The purpose of this methodological note is to provide a description of the processing of influential values in surveys. This document is split into five main parts. In the first part, we detail the theoretical framework and highlight the differences between outliers and influential values. In the second part, we discuss the reasons for the presence of influential values in surveys and present some good practices to adopt when creating surveys in order to limit the problem of influential values. The third part provides a presentation of a tool used for measuring influence : conditional bias. In the fourth part, we detail the methods used to handle the problem of influential values. Lastly, in the final part we provide some examples of the implementation of these methods by INSEE.**

## I. Theoretical Framework and Contexts for Application

### A. Context

distribution. For example, in the case of turnover, there is a positive variable of interest, the distribution of which is skewed to the right. In this context, there are often influential values present in the drawn sample. These are usually very large values, the presence of which in the sample tends to render traditional estimators (the expansion estimator, for example) very unstable. Robust estimators are created so as to limit the impact of the influential values, resulting in estimators that are more stable but potentially biased. The objective of robust estimation is to detect the influential values, then to develop robust estimation procedures which have a mean squared error significantly lower than traditional estimators in the presence of influential values. In addition, it is hoped that these estimators would not suffer from a significant loss of effectiveness in the absence of influential values. The processing of influential values therefore usually makes it possible to achieve a compromise between bias and variance.

### B. The Distinction between Outliers and Influential Values

It is important to make a distinction in our sample between two types of outlier units : representative outlier units and non-representative outlier units. This notion of representative units was introduced and discussed by Chambers (1986). The representative units, which will be considered potentially influential hereinafter, are units the collected value of which in the sample is correct and is not considered unique, in the sense that it is likely that there are other units in our U population with a collected value of the same order of magnitude. In the case of estimating a finite population parameter as a total, these units are of considerable importance in the estimation thereof and cannot be given a weight of 1, as that would be equivalent to considering them to be unique. The non-representative outliers are units the collected value of which is incorrect, due to a malfunction in the collection process : a classic case is the turnover of a company stated in euros instead of being stated in thousands of euros. The processing of this type of unit can be done at the data clearance stage, in particular through imputation processes : the turnover considered erroneous can be imputed using the turnover obtained from a previous survey. These outlier units are in fact unique and can be assigned a weight of 1 in the continuation of the estimation process or their value can be corrected if the error can be identified.

### C. Definition of the Notions of Configuration and Influential Value

Before defining the notion of influential value, the concept of configuration is introduced :

A configuration $\mathscr{C}$ is defined by the following quartet :

(1) a variable of interest $y$ ;
(2) a parameter of interest ;
(3) a sampling design ;
(4) an estimator.

The concept of configuration is a central notion insofar as a unit is influential in a given configuration ; i.e. a unit is influential for a given plan, parameter and estimator. In a given configuration $\mathscr{C}$ , a value will be defined as **influential** if it has a significant impact on the mean squared error of the estimator in question. An example is provided in Section (3.C) to illustrate the notions of configuration, influential unit and conditional bias.

## II. Why do we See Influential Values in our Surveys and How Can We Guard Against their Occurrence ?

### A. Variables of Interest with Asymmetric Distributions

In business surveys, it is common to see variables of interest, such as turnover, the distribution of which is highly asymmetric. Thus, certain businesses make a very significant contribution to the aggregate to be measured. Whether or not one of these “Big” units is selected has a significant impact on the estimator.

### B. Inter-Strata Migrants or “Strata Jumpers”

A second problem, resulting in the presence of influential values in the sample, is that of “strata jumpers”, which arise when the stratification information collected in the field is different from that available in the sampling frame. These differences are usually due to imperfections in the sampling frame (such as in the case of a slightly dated base, for example). A strata jumper is a unit that does not belong to the stratum to which it should have belonged if the information in the sampling frame was correct. If a unit with a high value is assigned to a non-exhaustive stratum, it will then combine a high value of the variable of interest and possibly a large survey weight, making it potentially very influential.

In practice, it is not rare to see between 5% and 10% of "strata jumpers". The older the sampling frame, the higher this percentage is.

### C. Poor Correlation between the Survey Weights and the Variable of Interest

It is possible to guard against the impact of influential values at the sampling method stage by automatically selecting potentially influential units. For example, in business surveys, it is customary to use a simple stratified random non-discount plan including one or more exhaustive strata, which are usually composed of large units. Unfortunately, it is rarely possible to completely eliminate the problem of influential values at the sampling method stage. Indeed, the strata in business surveys are usually formed by a size variable (for example, the number of employees as at 31 December of the previous year) and an activity classification variable (the APE code, for example). In a survey containing dozens of variables of interest, it is not unlikely that some of them have little or no link to the stratification variables, which can then lead to the presence of influential values.

In order to guard against the problem of influential values, it is also important to control the weight correction factors resulting from the adjustment methods :
- At the non-response correction stage, the use of re-weighting classes makes it possible to protect against extreme variations in the correction factors.
- At the calibration stage, it is possible to limit the weight variation by using an adapted distance function.

## III. How Can the Influence of a Unit be Quantified ?

### A. Conditional Bias : a Tool for Measuring Influence

The notion of influential value is relatively vague and a tool is needed to make it possible to measure influence while taking into account the sampling method. In the case of an approach under the plan, the notion of conditional bias has been developed in two articles by Moreno-Rebollo et al. (1995, 1999). This notion of conditional bias was used by Beaumont et al. (2013) in order to quantify the influence of a unit under the plan, so as to then create robust estimators.

Let $U = (1,...,k,...,N)$ be a finite population, $P(.)$ sampling design defined on $U$ and $Y$ is the variable of interest to be observed in the population. $\theta$ is the parameter of interest and $\hat{\theta}$ is an estimator of $\theta$. The conditional bias of a sampled unit $i$ associated with the estimator $\hat{\theta}$ is defined by : $B_i^{\hat{\theta}}(I_i = 1) = \mathbb{E}_P\left(\hat{\theta}|I_i = 1\right) - \mathbb{E}_P\left(\hat{\theta}\right)$, where $I_i$ is the binary variable indicating belonging to the sample which takes the value 1 if the unit $i$ is in the sample, or 0 otherwise. Similarly, the conditional bias of an unsampled unit $i$ associated with the estimator $\hat{\theta}$ is defined by :

$$B_i^{\hat{\theta}}(I_i = 0) = \mathbb{E}_P\left(\hat{\theta}|I_i = 0\right) - \mathbb{E}_P\left(\hat{\theta}\right).$$

Conditional bias is a measurement of an influence as it makes it possible to observe the average impact caused to the estimator, depending on whether or not the unit $i$ belongs to the sample. It is important to note that the conditional bias of an unsampled unit is unknown, and it is impossible to estimate because that involves values for the variable of interest outside of the sample. Therefore, there is no protection against the influence of unsampled units.

We can explicitly compute the influence on the Horvitz-Thompson estimator defined by :

$$\hat{t}_{y\pi} = \sum_{j\in S} \frac{y_j}{\pi_j}.$$

where $\pi_j$ is the probability of inclusion of the unit $j$. The formula $d_j = \frac{1}{\pi_j}$ identifies the survey weight of the unit $j$. The conditional bias of a sampled unit $i$ for the Horvitz-Thompson estimator is defined by :

$$\begin{aligned} B_i^{HT}(I_i = 1) &= E_p\left(\hat{t}_{y\pi}|I_i = 1\right) - t_y \\ &= \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) y_j. \end{aligned} \quad (1)$$

Given that the expression of conditional bias involves first- and second-order inclusion probabilities, it takes account of the sampling method.

### B. The Properties of Conditional Bias

1) A unit with an inclusion probability equal to 1 is found to have a conditional bias of zero, i.e. the influence of that unit is equal to zero. Thus, it is understood here that the use of an exhaustive stratum makes perfect sense.

2) It is also important to note that the sampling error of the Horvitz-Thompson estimator $\hat{t}_{y\pi} - t_y$ can be broken down as follows

$$\hat{t}_{y\pi} - t_y = \sum_{i\in S} B_i^{HT}(I_i = 1) + \sum_{i\in U\setminus S} B_i^{HT}(I_i = 0) \quad (2)$$

if

$$\sum_{i\in U}(I_i - \pi_i)\, a_i = 0, \quad (3)$$

where $a_i = (1-\pi_i)^{-1}\left\{B_i^{HT}(I_i = 1) - (d_i - 1)\, y_i\right\}$.
It can be shown that the condition (3) is verified for a Poisson sampling method. The breakdown (2) is approximately respected for a stratified random sampling method without replacement or a high entropy fixed-size sampling method. In the event that the breakdown (2) is valid, the conditional bias can be seen as the contribution of the unit $i$ to the sampling error $\hat{t}_{y\pi} - t_y$.

3) The following property holds for any survey sampling design $P(.)$ :

$$Var_p(\hat{t}_{y\pi}) = \sum_{j\in U}\sum_{k\in U}\frac{y_j y_k}{\pi_j\pi_k}\Delta_{jk} = \sum_{i\in U} B_i^{HT}(I_i = 1) y_i. \quad (4)$$

where $\Delta_{jk}$ is the variance-covariance matrix of the indicators $I_j$ and $I_k$.

The variance of the Horvitz-Thompson estimator is therefore directly related to the conditional bias, and it is found that a unit with a strong conditional bias will contribute to the variance significantly. In addition, the higher the value of the variable of interest $y_i$ the greater its contribution to the variance.

### C. Example for Two Specific Sampling Methods

Let us consider a population size of 5000,for which we observe fictional turnovers in thousands of euros $y$, sorted by

ascending order :

$$y_1 = 0, y_2 = 500, y_3 = ... = y_{4999} = 500 \text{ et } y_{5000} = 2000$$

In this case, the average in the population $\bar{y}_U$ is equal to 500.2. Let us assume that we are in one of the two following configurations :

$\mathscr{C}_1$ : (Turnover, Total turnover, Simple random sampling without replacement, Horvitz-Thompson estimator)

$\mathscr{C}_2$ : (Turnover, Total turnover, Poisson sampling with equal probabilities $\pi_k = \frac{n}{N}, k \in U$, Horvitz-Thompson estimator)

In order to make the link between the conditional bias and the instability of the estimators, in Table 1 we reiterate the conditional bias associated with a selected unit and the variance formulas for the Horvitz-Thompson estimator.

| | Variance | Unit $i$ conditional bias |
|---|---|---|
| Simple Random Sampling without replacement | $N^2 \frac{(1-\frac{n}{N})}{n} S^2_{y_U}$ | $\frac{N}{N-1}(\frac{N}{n}-1)(y_i - \bar{y}_U)$ |
| Poisson sampling | $\sum_{k \in U} \frac{(1-\pi_k)y_k^2}{\pi_k}$ | $(d_i - 1)y_i$ |

Tableau 1 : Summary of the Variance Formulas and Conditional Bias for the Horvitz-Thompson Estimator

In the case of simple random sampling without replacement, the first unit with a turnover equal to 0 contributes significantly to the variance for the Horvitz-Thompson estimator if selected (it contributes to a high value for dispersion $S^2_{y_U}$), while in the case of Poisson sampling, the first unit does not contribute to the variance of the Horvitz-Thompson estimator (the term $k = 1$ is zero in the variance formula associated with Poisson Sampling). Thus, the influence of the unit is highly dependant on the sampling design used. This can be seen directly for each unit using conditional bias : in the first case, the conditional bias is very high as the value 0 is very far from the average $\bar{y}_U = 500,2$. Whereas in the case of Poisson sampling with equal probabilities $\pi_k = \frac{n}{N}, k \in U$, the conditional bias is zero, since $y_1 = 0$ and therefore the influence of the first unit is zero in the second configuration, whereas it is high in the first configuration. Finally, a unit with a value of $y_{5000} = 2000$, is influential for both sampling designs.

## IV. How Should the Problem of Influential Values be Handled ?

### A. The Traditional Winsorisation Method

In practice, one method that is used in particular is winsorisation, which consists of reducing the sample values that are too high to a certain threshold. In the case of winsorisation, a unit is considered to have an influence if the product of its weight and its value exceeds a certain threshold. The literature distinguishes between two types of winsorisation. Standard winsorisation, also known as type 1 winsorisation in the case of simple random sampling without replacement, consists of reducing the value of units that exceed a certain threshold, taking into account their weight. So, $\tilde{y}_i$ is the value of the variable $y$ for the unit $i$ after winsorisation. This gives

$$\tilde{y}_i = \begin{cases} y_i & \text{si } d_i y_i \leq K \\ \frac{K}{d_i} & \text{si } d_i y_i > K \end{cases} \quad (5)$$

where $K > 0$ is the winsorisation threshold. The standard winsorised estimator of the total $t_y$ is defined by

$$\hat{t}_s = \sum_{i \in S} d_i \tilde{y}_i \quad (6)$$

Another way of writing it entails expressing $\hat{t}_s$ as a weighted sum of the initial values using modified weights :

$$\hat{t}_s = \sum_{i \in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = d_i \frac{min\left(y_i, \frac{K}{d_i}\right)}{y_i}. \quad (7)$$

If $min\left(y_i, \frac{K}{d_i}\right) = y_i$ (i.e. the unit i is not influential), then $\tilde{d}_i = d_i$. The weight of a non-influential unit is therefore not amended. In contrast, the amended weight of an influential unit is less than $d_i$ and may even be less than à 1. It should be noted that a unit displaying a value $y_i = 0$ poses no particular problem since its contribution to the estimated total, $\hat{t}_s$, is zero. In this case, an arbitrary value can be assigned to the amended weight $\tilde{d}_i$. From a practical point of view, it is inconvenient to assign a weight of less than 1 to a unit, as we want it to at least be represented. This is why Dalén-Tambay winsorisation, also known as type 2 winsorisation, is generally preferred in the case of simple random sampling without replacement. The values of the variable of interest are defined after winsorisation by

$$\tilde{y}_i = \begin{cases} y_i & \text{si } d_i y_i \leq K \\ \frac{K}{d_i} + \frac{1}{d_i}(y_i - \frac{K}{d_i}) & \text{si } d_i y_i > K \end{cases} \quad (8)$$

This leads to the winsorised estimator of the total $t_y$ :

$$\hat{t}_{DT} = \sum_{i \in S} d_i \tilde{y}_i. \quad (9)$$

As for $\hat{t}_s$, an alternative way of writing it entails expressing $\hat{t}_{DT}$ as a weighted sum of the initial values using amended weights :

$$\hat{t}_{DT} = \sum_{i \in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = 1 + (d_i - 1)\frac{min\left(y_i, \frac{K}{d_i}\right)}{y_i}. \quad (10)$$

As for the standard winsorised estimator, the weight of a non-influential unit is not amended. Once again, a unit displaying a value $y_i = 0$ poses no particular problem since its contribution to the estimated total, $\hat{t}_{DT}$, is zero. In this case, an arbitrary value can be assigned to the amended weight $\tilde{d}_i$.

### B. Selecting the Threshold : a Decisive Choice in the Bias-Variance Trade-Off

Virtually all robust methods involve the use of a threshold. Selecting the threshold $K$ is very important as it makes it possible to make the bias-variance trade-off for the robust estimator. In practice, there are three ways to select this threshold :

- based on an expert statement : this choice is extremely risky because it can generate a robust estimator that is less efficient in terms of squared error than the initial non-robust estimator.

- by minimising the estimated mean squared error of the robust estimator ; here, for example, the method of Kokic and Bell (1994) can be cited, which applies in the case of estimation of the total for a positive variable of interest based on a sample selected by simple, stratified, single-stage random sampling, assuming that, in each stratum, the values of the variable of interest are created based on a single law of probability and that we have observations of the variable of interest in each independent stratum of the sample (for example, resulting from the sampling frame or a previous survey).
- by choosing the threshold that most minimises the calculated influences on the robust estimator : the details of this method are provided in the article by Beaumont et al. (2013).

In accordance with the data available in the sampling frames or in previous surveys, and the complexity of the sampling method, it may be necessary to use point 2 or 3. The minimisation of the mean squared error of the robust estimator is relatively complex and its implementation is only feasible for simple parameters such as total or average and for fairly simple plans : simple random stratified or Poisson sampling. Furthermore, it is very difficult to generally apply this method based on the minimisation of the mean squared error to account for non-response modelling as well as the calibration stage. The methods based on conditional bias make it possible to take these two essential stages into account in the adjustment of a survey. For further details on the general application of methods based on conditional bias that make it possible to take account of the non-response phase and calibration adjustments, readers may refer to the articles by Favre-Martinoz et al. (2015, 2016).

## V. An Example of the Processing of Influential Values : the Case of the ESANE Survey

Since 2008, the surveys of the ESANE (Élaboration des Statistiques Annuelles d'Entreprises - Elaboration of annual statistics of companies) scheme use winsorisation techniques in accordance with the method proposed by Kokic and Bell (1994). This method assumes the availability of data, from outside of the survey, on the distribution of the winsorised variable in the sampling strata. The ESANE scheme makes it possible to have the fiscal turnover of all companies in the sampling frame to define the winsorisation thresholds that are used, from 2013 onwards, for winsorisation. Once the winsorisation thresholds have been determined for the turnover variable, the question of how to process the other variables of company tax returns is asked. We would like to reiterate that company tax returns contain a great number of variables, which are linked together by numerous accounting relationships. A company may be atypical for only some of these variables. In addition, it would be possible to carry out a separate winsorisation for each variable of the tax return. Thresholds would be calculated for turnover, added value, gross operating surplus, investment, etc., and the atypical values would be identified and processed based on those thresholds. However, this method risks breaking the accounting relationships that exist between the variables in a single return for the winsorised units. A reasoned choice would be to calculate the winsorised weights corresponding to the thresholds determined using the Kokic and Bell (1994) method and use those weights for the other variables of interest of the survey. This adjustment is effective if the other variables of interest are closely correlated with turnover. This is the case, for example, in respect of added valued and payroll. In contrast, this method can be problematic when it comes to detecting influential values with little correlation with turnover, such as investment which, generally, is a complicated variable to process that presents a high level of variance and low temporal consistency.

## References


[1] Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555–569.

[2] Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063–1069.

[3] Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.

[4] Deroyon, T. (2015). Traitement des valeurs atypiques d'une enquête par winsorization. Application aux enquêtes sectorielles annuelles. Acte des Journées de Méthodologie Statistique de l'Insee, 1$^{er}$ avril 2015, Paris.

[5] Favre-Martinoz, C., Beaumont, J.-F., Haziza, D.(2015) *Une méthode de détermination du seuil pour la winsorisation avec application à l'estimation pour des domaines*, Techniques d'enquête, 2015.

[6] Favre-Martinoz, C., Haziza, D., Beaumont, J.-F.(2016). *Robust Inference in Two-phase Sampling Designs with Application to Unit Nonresponse*. Scandinavian Journal of Statistics.

[7] Kokic, P.N. and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419–435.

[8] Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling : conditional bias. *Biometrika*, 86, 923–928.

[9] Munoz-Pichardo, J., Munoz-Garcia, J., Moreno-Rebollo, J. and Pino-Mejias, R. (1995). A new approach to influence analysis in linear models. *Sankhyā : The Indian Journal of Statistics, Series A*, 393–409.




*Département des méthodes statistiques*
*Version n$^{o}$ 1, diffusée le 10 octobre 2017*