

ON WINSORISATION IN BUSINESS SURVEYS

Philip Kocio¹

ABSTRACT

The use of sample surveys to provide population estimates for variables in finite populations is widespread, but can be crucially affected by outliers. The method of winsorisation, which can be used for the detection and treatment of the effects of outlying observations with respect to an estimation model, is extended to detect and treat observations which are unusually large or unusually small, instead of the more common treatment in sample surveys of only large outliers. A method is presented to determine the appropriate boundaries for outlying observations, based on previous survey data. Some simulation results based on UK retail sales inquiry data are presented, which demonstrate that the new method has better properties than one-sided winsorisation.

KEY WORDS: Winsorization; Outlier; Sample surveys; Generalised regression estimation.

RÉSUMÉ

L'utilisation des enquêtes par sondage pour fournir des estimations pour des populations finies est très répandue, mais peut être affecté de façon cruciale par des unités aberrantes. La méthode de Winsorisation, qui peut être utilisée pour la détection et le traitement des effets des observations aberrantes par rapport à un modèle d'estimation, est étendue à la détection et au traitement des observations qui sont exceptionnellement grandes ou exceptionnellement petites, plutôt que du traitement plus courant dans les enquêtes par sondage pour seulement les grandes unités aberrantes. On présente une méthode pour déterminer les bornes appropriées pour les observations aberrantes, basé sur les données de l'enquête précédente. On présente quelques résultats de simulation basés sur les données de l'enquête des ventes au détail au Royaume-Uni, ce qui démontre que la nouvelle méthode a de meilleures propriétés que la winsorisation à un seul côté.

MOTS-CLÉS: Winsorisation; données aberrantes; enquêtes par sondage; estimateur par régression généralisé.

1. INTRODUCTION

Outliers in surveys arise from a number of sources and can have a large effect on estimates of totals. This is especially true in business surveys where the populations from which the samples are drawn are typically skewed, but with outliers in both directions. We will consider here only *representative outliers* (Chambers, 1986), that is extreme observations that have been confirmed by the contributor, and consequently under our random sampling assumptions represent similar unsampled observations in the population.

Winsorisation is a value-modification method originally applied to sample surveys by Searls (1966). It is a method of outlier treatment and it has typically been applied to unusually large observations (Kocio and Bell, 1994; Rivest and Hurtubise, 1995). We shall refer to these kind of winsorisation methods as one-sided. Clarke

(1995) extended the method to a general class of estimators of total including the GREG estimator (Särndal *et al.*, 1992). The method operates by replace survey values, y_i , greater than an observation-specific bound, K_i , by $\{y_i + (w_i - 1)K_i\}/w_i$, where w_i is the survey weight. The survey value is left unmodified if it is less than this bound. Denote the winsorised value by y_i^* and the resulting one-sided winsorised estimator of total by $\hat{T}^* = \sum_{i \in s} w_i y_i^*$, where s are the sampled units. The optimal bounds are obtained with $K_i = \hat{\mu}_i + M/(w_i - 1)$, where for member i , $\hat{\mu}_i$ is the fitted value under the assumed estimation model. The value M can be estimated from previous survey data by following a specific procedure which balances bias against variance; see Cruddas and Kocio (1996) for details. The main practical disadvantage with one-sided winsorising is that the resulting estimator has negative

¹ Philip Kocio, Insider Teknon GmbH, Wilhelm-Theodor-Römhheld-Straße 32, Mainz, Germany, D-55130, ko@tekon.de

bias, which makes it problematic for general use in surveys where derived variables may, as a result, become unusable due to the aggregation of biases across several winsorised survey variables.

In the following section we briefly describe a two-sided procedure that partly overcomes the shortcoming above. In section 3 the results of a simulation study are presented, and finally in section 4 these results are discussed.

2. THE TWO-SIDED PROCEDURE

In the two-sided winsorising procedure the survey value is adjusted if it is either too large or too small with respect to the estimation model. The modified y -values are constructed as follows. Let

$$y'_i = \begin{cases} \hat{\mu}_i + U\hat{\sigma}_i, & \text{if } y_i > \hat{\mu}_i + U\hat{\sigma}_i, \\ \hat{\mu}_i - L\hat{\sigma}_i, & \text{if } y_i < \hat{\mu}_i - L\hat{\sigma}_i, \\ y_i, & \text{otherwise,} \end{cases}$$

where $U, L > 0$ and $\hat{\sigma}_i$ is the predicted value of scale under the working model which is usually based on some robust estimate of dispersion. Following the type II winsorisation approach suggested by Gross *et al.* (1986), the winsorised y -values are $y_i^{**} = \{y_i + (w_i - 1)y'_i\}/w_i$ and the two-sided winsorised estimator of total is $\hat{T}^{**} = \sum_{i \in N} w_i y_i^{**}$. Note that this reduces to a specific type of one-sided winsorisation as the lower cut-off parameter $L \rightarrow \infty$.

The motivation for this procedure is two-fold. Firstly, in the symmetric case when $U = L$, \hat{T}^{**} can be viewed as a one-step M -estimator of total (Chamber and Kocic, 1993). Secondly, in theory to be presented in a forthcoming paper, it is shown for the ratio estimator that if the standardised residual errors under the working model are independent with common distribution F and density function f , and if a statistically independent but identically distributed estimate of location is used in place of $\hat{\mu}_i$, then the optimal two-sided bounds approximately satisfy

$$F(-L)dL = \{1 - F(U)\}dU \text{ and}$$

$$U + L = \left(1 + \frac{dU}{dL}\right) \left\{f(U) + f(-L) \frac{dL}{dU}\right\}.$$

Under these conditions the winsorised ratio estimator will be approximately unbiased.

For many cases including the standardised normal and lognormal distributions, the solution to the differential equations above satisfy $U + L \leq 3$. This suggests that the following procedure can be used for estimating the two-sided parameters from historic data: (i) initially set $U = L = 1.5$; (ii) estimate the bias of \hat{T}^{**} by comparing it with the non-winsorised estimator \hat{T} ; (iii) if the bias is negative, then add a small amount, $\Delta = 10^{-3}$ say, to U and subtract Δ from L , and if the bias is positive subtract Δ from U and add it to L ; (iv) repeat from step (ii), possibly modifying the value of Δ until the estimated relative bias of \hat{T}^{**} is small. This procedure can be applied at several points in time when repeated sample survey data is available and then the parameter estimates can be averaged to produce more accurate results.

3. SIMULATION

A simulation study involving real data from the UK retail sales inquiry (RSI) has been undertaken to investigate the effects of using winsorisation. The survey variable y is average weekly retail sales, and the auxiliary variable x is register turnover. Six consecutive months of RSI survey data was available. Data from the first 3 months was pooled and treated as a single (design) population for estimating the cut-off parameters and for sample design purposes, while data from the last three months was used to assess the various estimators. Each of the two populations were stratified into 6 broad industry groups by 5 size strata. The largest size strata were completely enumerated and a five per cent sample was allocated, using the Neyman allocation rule, to the remaining strata. In each simulation run five independent samples were drawn from the design population and separate winsorising parameters were estimated in each broad industry group using the procedure outlined in section 2. An independent sample was then drawn from the assessment population and winsorised and non-winsorised ratio estimates were computed. A total of 1000 simulations were performed and various statistics (as in table 1) were computed to show the difference between the true population value and the estimated values.

Table 1. Estimates of bias, root mean squared error and coverage probabilities of 95 per cent confidence intervals based on 1000 independent simulations from the RSI data.

Industry	Bias (per cent of total)			RMSE (per cent of true total)			Coverage probability (percent)		
	No treatment	1-sided winsor.	2-sided winsor.	No treatment	1-sided winsor.	2-sided winsor.	No treatment	1-sided winsor.	2-sided winsor.
All	-0.01	-0.71	0.00	1.29	1.09	0.98	91.2	84.4	94.8
521	0.01	-0.25	-0.15	0.49	0.45	0.42	88.0	84.4	86.8
522	-0.10	-1.34	-0.79	4.74	3.65	3.84	90.0	89.2	92.4
523	0.20	-1.83	0.13	4.95	4.39	4.85	90.0	86.4	90.8
524	-0.15	-1.11	0.20	3.46	2.45	2.61	88.8	88.0	94.0
525	-0.79	-6.94	-2.35	16.29	11.68	14.29	77.2	69.6	79.2
526-7	0.68	-1.66	0.67	7.40	5.55	6.72	84.8	84.4	90.4

4. DISCUSSION

From the results above it can be seen that the bias of the one-sided winsorising is quite large, which may be of concern in certain practical situations, whereas the bias of the two-sided winsorised estimator is considerably smaller, although in general not always as small as the ordinary ratio estimator. Furthermore, unlike the one-sided version, the two-sided winsorised estimator does not appear to have a consistent direction to its bias, which is preferable when forming derived variables or when aggregating estimates across domains.

The root mean squared error (RMSE) of both winsorised estimators is considerably less than in the no treatment case. At the all industries level the two-sided winsorised estimator is about 10 per cent more efficient than the one-sided winsorised estimator, and about 30 per cent more efficient than the ordinary ratio estimator.

Finally, the coverage probabilities of the confidence intervals are in general closest to the nominal value of 95 per cent in the case of the two-sided winsorised estimator and, due to the significant impact of bias, are quite poor for the one-sided winsorised estimator.

REFERENCES

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

Chambers, R.L. and Kokic, P.N. (1993). Outlier robust sample survey inference. Proceedings of the ISI 49th Session, Firenze, Italy.

Clark, R.G. (1995). Winsorisation methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.

Cruddas, M., and Kokic, P.N. (1996). The treatment of outliers in ONS business surveys. Proceedings of the GSS(M) methodology conference. Newport, UK.

Gross, W. F., Bode, G., Taylor, J.M., and Lloyd-Smith, C.W. (1986). Some finite population estimators which reduce the contribution of outliers. Proceedings of the Pacific Statistical Congress, Auckland, New Zealand, 20-24 May 1985.

Kokic, P., and Bell, P.A. (1994). Optimal winsorizing cut-offs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419-435.

Rivest, L.-P., and Hurtubise, D. (1995). On Searls' winsorised mean for skewed populations. *Survey Methodology*, 21, 107-116.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Searls, D.T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.