

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 6 of the provisional agenda

WINSORISATION: AN UPDATE

Submitted by the Office for National Statistics of the United Kingdom

¹ Prepared by Paul Smith.

Abstract

An update on the use of winsorisation within ONS business surveys is given, including further information on the practical uses and problems associated with the technique. Links with the editing and validation process are drawn, and some suggestions made for the focus of further research. Some methods for employing winsorisation in sampling frames, and some initial results from using winsorisation techniques to detect frame anomalies are presented.

I. INTRODUCTION

1. Winsorisation is a technique for reducing the effect of outlying observations on survey estimation. The details are given in Smith & Kokic (1996). Basically, a minimum mean squared error criterion is used in an algorithm to construct a parameter which determines whether an observation lies a long way away from the model. Since both the value of the observation and its weight in the estimation process are important, the decision rule on whether an observation is an outlier, and the extent to which its value (and hence influence on the estimates) is reduced, depend on both these factors.

2. Outliers are observations from a dataset which are extreme in comparison with the main part of the data. They can be divided into two types, those which arise from errors in the way the data are collected, transferred and processed, and those which are essentially accurate, but which are a long way away from the other data. For example, if a business fills in a form with its sales in £th when the question asks for the response in £m, there will clearly be a data error, and (following Chambers 1986) these will be called *unrepresentative outliers*, because they do not contribute to the sampling variability through their representation of non-sampled units². They are essentially errors to be (where possible) identified and corrected. This is the normal scope of the editing processes.

3. The second type of outlier is an observation which is correct, but which lies away from the main body of the data. This is not an error, and the outlying observation will, in the random sampling sense, represent some non-sampled observations. That is, under the assumptions of random sampling, there will be some population elements with similar characteristics to this sample element, even though only one such element has been observed. These are *representative outliers* (Chambers 1986). This sort of outlier is illustrated diagrammatically in Figure 1.

4. How to deal with representative outliers is a difficult problem, mainly because outliers may be finite in number, and occur rarely. Most (non-survey) sampling theory is based on the idea of an infinite population, so that there is an infinite number of each type of element. In survey sampling, however, we typically have a finite population, and only sample some of it. As an example, 1 in 10 businesses from a population of 200 businesses might be sampled, so each sample member represents itself and 9 non-sampled businesses. The difficulty arises if it is suspected that there are say, only two elements with particular

² Unless we are prepared to make the very specific assumption that the process which gives rise to these errors is such that they occur randomly.

characteristics (“outliers”); if one is picked it will be assumed by default that it represents nine more even though there aren’t that many in the population, and if one is not picked then the outlying units will be assumed to be just like all the sampled ones.

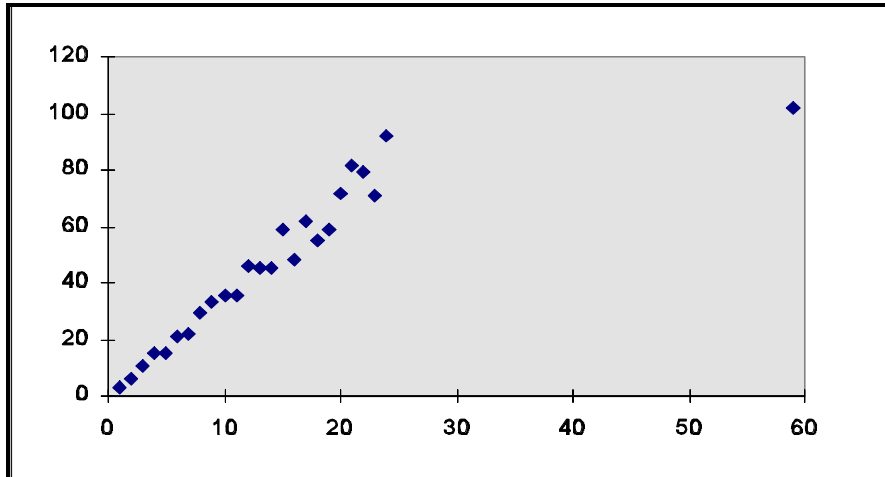


Figure 1 : Diagrammatic illustration of a representative outlier. The observation on the extreme right is correct, but unusual (“outlying”) compared with the other observations.

5. There are restrictions in what can be done by what information is available. It is not known how many outliers there are in the population, so the sample weighting cannot be modified to take account of the number of outliers, although approximations can be made (Hidioglou & Srinath 1981). In the case where no outliers are detected, there is no information on which to base estimates for outlying observations; it is not even known if there are any! In this case, the problem would typically be ignored. In a repeating survey we might try to get some idea of what proportion of businesses is outlying over a long period, and use this to improve the stability of our estimation. Winsorisation is a method which is designed to treat representative outliers. Note that these outliers are only defined with respect to a model - they don’t deviate from an expected value unless you have a model (explicit or implicit) which says what the value “ought” to be.

II. WINSORISATION

6. The theory of winsorisation currently deals with a population where the returned values are heavily skewed to the left, that is with a long right tail to the distribution so that there are occasional large values. This distribution is of the residuals compared with the estimation model being used. This means (a) that only the largest returned values are affected (small values cannot be outlying), and, more or less equivalently, (b) that extreme values in the auxiliary variable(s) are not modified in any way. That is, there is an implicit assumption that the auxiliary variables are correct.

7. The form of the threshold defined by the winsorisation process for various estimation models was shown in Smith & Kocic (1996). For number raised estimation, it is $y = \mu + K$ for some constant K where μ is the expected value for the observation under the estimation model. That is, any observation bigger than $\mu + K$ is an outlier. For ratio estimation, the threshold is a line parallel to the fitted ratio line, but with an intercept $\gg 0$.

8. The general theory for regression estimation (and more complex calibration estimators) follows the same derivation (Cruddas & Kokic 1996). At this stage, however, practical considerations begin to apply. Regression estimators can in general lead to observations having negative weights. Although the theory applies equally to these cases, the algorithm for determining the optimal cut-off parameter does not work properly where there are negative weights. The possible work-arounds for this are:

(i) to produce a set of calibrated weights with additional constraints to ensure that all weights > 0 ; use these both for parameter determination and winsorisation/estimation;

(ii) produce a set of weights with constraints as in (i), and use this to calculate the threshold parameter. Then use the unconstrained weights in the estimation and winsorisation process.

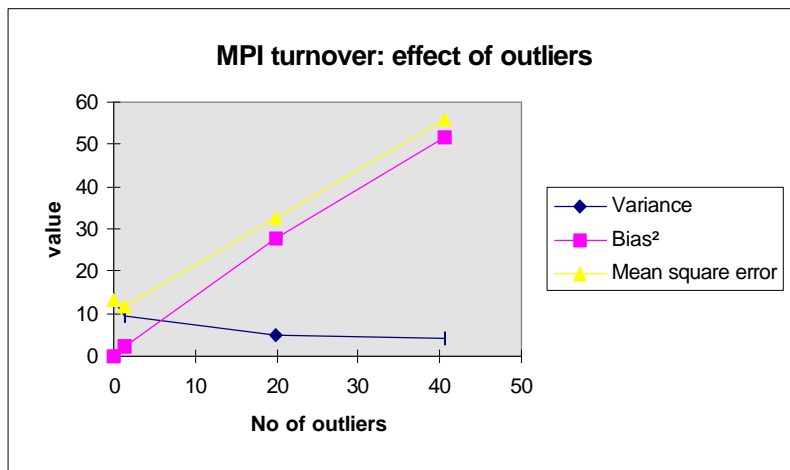


Figure 2: Contributions of variance and bias to mean squared error with differing numbers of outliers, for turnover values from the Monthly Production Inquiry. Plotted points are averages over 12 months.

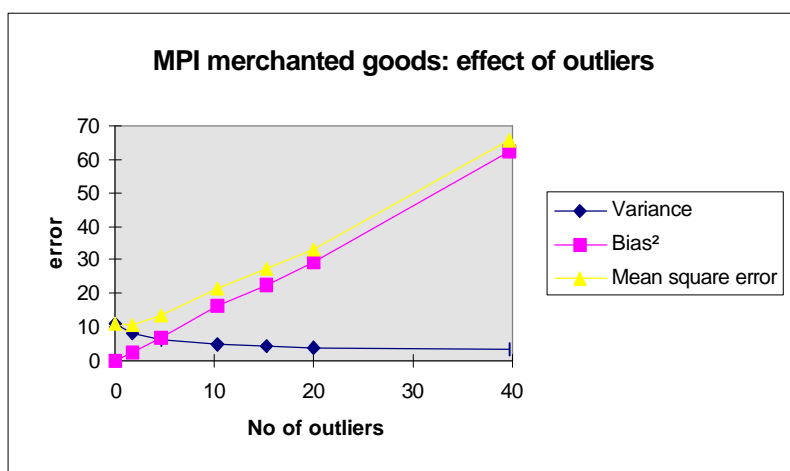


Figure 3: Contributions of variance and bias to mean squared error with differing numbers of outliers, for merchanted goods values from the Monthly Production Inquiry. Plotted points are averages over 12 months.

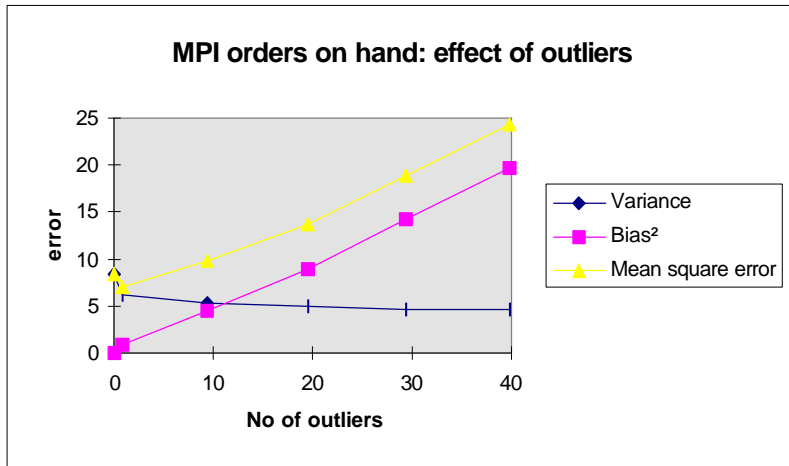


Figure 4: Contributions of variance and bias to mean squared error with differing numbers of outliers, for orders on hand values from the Monthly Production Inquiry. Plotted points are averages over 12 months.

III. RESULTS OF PRACTICAL APPLICATION

9. When these methods are applied to data from the Monthly Production Inquiry (MPI) (a monthly survey of output of manufacturing industries in Britain), some surprising results are obtained - the optimum value of the mean square error at the whole-survey level is obtained when very few observations are treated as outliers. This is shown in Figures 2 to 4, depicting the results from three different variables collected in this survey. For the main variable, turnover, only 1.7 outliers per month on average gives the optimum mse properties of the estimator. If many more outliers are marked (more in line with practice when winsorisation is not used), it is clear that a large bias into the estimation is introduced, in order to make the variance small.

10. One of the implicit assumptions in winsorisation is that there is a clean data set with no errors before starting. This is especially important when the threshold parameter is being calculated. What is actually happening is that the editing process is leaky, and abnormal observations are contributing in the calculation of the threshold parameter. These observations typically have the largest effect on the estimate (they are wrong, not just unusual), and so treating them gives the best mse properties. However, they do mask the effect of the real outliers, which are then not properly identified.

11. Since the assumption that having a clean dataset is one we would not be willing to make (in fact following up the most extreme cases from the threshold analysis has identified some frame errors and some measurement errors), the possibility of using winsorisation as a method of identifying and treating outliers is being investigated, but setting the threshold parameter to give an apparently non-optimal MSE by marking many more outliers. For MPI some tests using 40 outliers are being done per month as an average (which is still fewer than on the survey before introducing winsorisation). From Figure 5, it can be seen that this is the point at which the change in estimates from month to month becomes more stable, and after which many more outliers have to be marked to make any further difference. Looking at

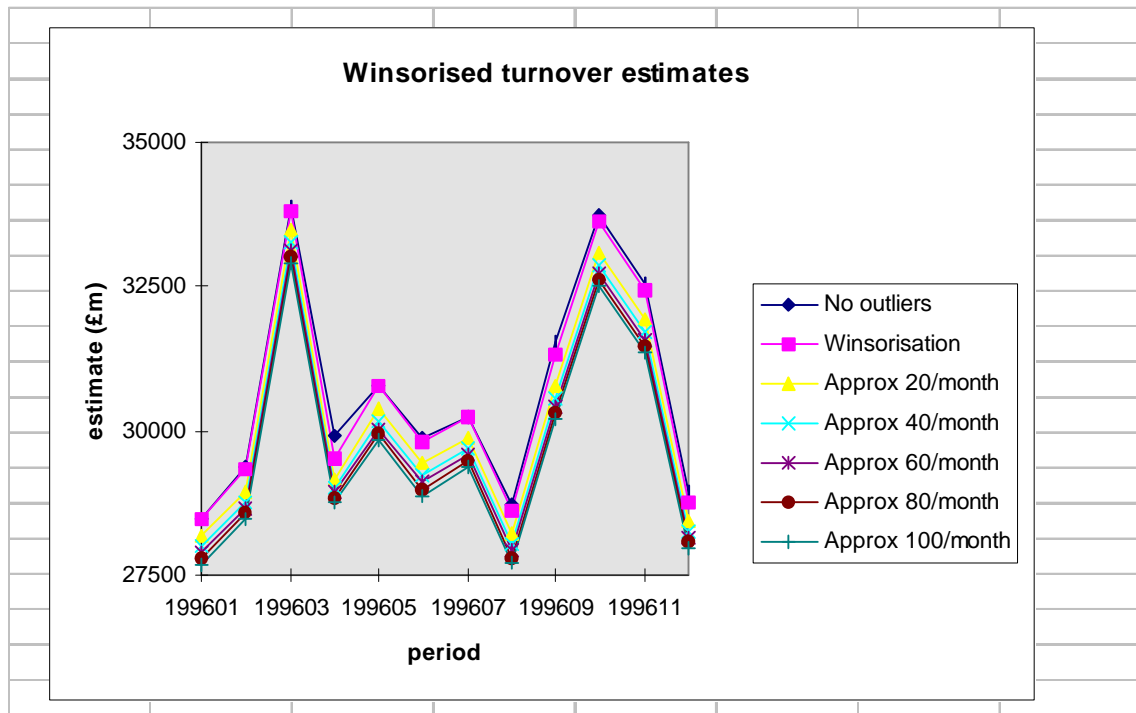


Figure 5: Whole survey estimates of turnover from the Monthly Production Inquiry under scenarios with a variety of numbers of outliers per month.

Figure 2, however, we are a long way from the optimum MSE, and can only justify this (*post-hoc*) by saying that some of the bias is correcting for measurement error, and is not contributing to the estimation MSE.

IV. WINSORISATION IN FRAMES

12. The discussion above suggests that winsorisation may be identifying errors which should be picked up by the validation process. This leads to the idea that winsorisation can be used as part of the data validation toolkit. One unusual way this idea can be applied is to see which observations on our sampling frame (the Inter-departmental business register (IDBR) for business surveys in ONS) are the most extreme with respect to a common estimation model. There is one main problem to be overcome. It was noted that winsorisation finds those observations with the largest contribution to an estimate. This contribution is to do with their *representativity*, as shown by their weight. When the weight is 1.00, this means that the element has been sampled with certainty, and that it represents only itself. It has no contribution to the estimate through representing other units, and hence cannot be an outlier.

13. On a frame where an estimate is not made, and a sample is not done - the (auxiliary) values for the whole population are known. However, if each member of the

population is given a weight of 1.00, there are no outliers. So it is necessary to generate some “typical” weights for these elements in ONS business surveys. This is done by dividing the register into strata (based on industrial sector and employment size), finding the total survey sample in a stratum for all ONS surveys, and dividing this number by the population size. (It might be better to use the number of businesses which are selected for any survey, divided by the population size.) This gives an approximation to the weight an observation might get if it was selected in any survey. This weight is then used to calculate the appropriate threshold parameter, and to determine the outliers on the IDBR.

14. There are two main auxiliary variables on the IDBR, turnover and employment. Using expansion estimation and the turnover auxiliary variable as the “survey” variable, only one outlier is found, with a turnover £27bn and an approximate weight of 70, so its contribution if selected in a survey would be £1890bn. This one is so extreme that it masks any other outliers/errors on the frame.

15. Using employment, there are fewer distinct values in each stratum, so there are 227 outliers, all in the same stratum, and all with the same characteristics - employment is 4, approximate weight is 39. Removing this stratum from consideration puts all the outliers in a different stratum, and so on. This clearly demonstrates the masking effect - removing the worst case puts all the power of the outlier identification on the next worst case, and this procedure can be continued through several stages.

16. The proof of the accuracy of the register may be to compare employment and turnover using a ratio-type estimator (most estimation in ONS business surveys is based on the relationship between these two variables measured in some way or another). It turns out that the initial turnover outlier masks everything else again... its employment is 4! And that means it is a very long way from the fitted ratio line for its stratum.

17. Despite this lack of identification of extreme observations in accordance with the “minimum MSE estimator” principle, there is still merit in using winsorisation in frames. By setting the threshold parameter to a much lower value than the “optimal” one from the MSE calculations, many more observations can be marked as extreme. There is additional gain in that the proportion by which the value of the observations would be reduced if using this method in the estimation process gives an ordering of the observations which deviate most from the assumed estimation model. This gives a clear indication of where resources would be most profitably directed to check that the frame provides a sound base for sampling.

V. CONCLUSIONS

18. Using winsorisation in practice suggests that there should be many fewer outliers than have previously been used in producing estimates. Much of this may be due to the presence of measurement errors in the data, but the techniques of winsorisation can be used to identify, and even treat, these, as long as certain assumptions can be made about the data quality (specifically, the likely number of errors in the survey). Winsorisation may provide new leads in the field of data editing by providing an objective criterion for working out which edit failures to follow up first, depending on their influence on the final

estimate.

References

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

CRUDDAS, M. & KOKIC, P. (1996) The treatment of outliers in ONS business surveys. *Proceedings of the GSS(M) Methodology Conference 25 June 1996*.

HIDIROGLOU, M.A. & SRINATH, K.P. (1981) Some estimators of a population total containing large units. *Journal of the American Statistical Association* **78** 690-695.

SMITH, P.A. & KOKIC, P. (1996) Winsorisation in ONS business surveys. Working paper no. 22 at the UN Data Editing Conference 1996, Voorburg.