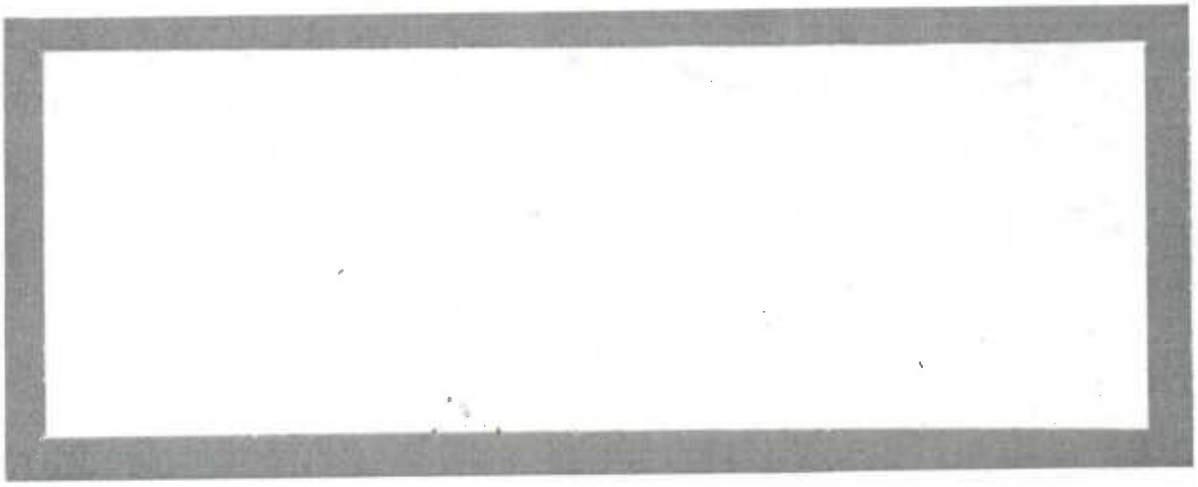

Methodology Branch

Institutions & Agriculture
Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquête
institutions et agriculture

Canada

WORKING PAPER No. IASM-85-056E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL No. DMEIA

MÉTHODOLOGIE

OUTLIERS & INFLUENTIAL OBSERVATIONS IN SAMPLE SURVEYS

by

Jack Gambino

March, 1985

* This is a preliminary version. Do not quote without author's permission. Comments are welcome

Outliers and Influential Observations
in Sample Surveys

ABSTRACT

Sample survey data sets often contain one or a few observations which have a much larger impact on estimates than seems reasonable. This report studies methods of dealing with such data, with emphasis on area frame problems. The National Farm Survey is used to illustrate both the problems and some possible solutions.

Observations Abérrantes et Observations Influentes
dans les Enquêtes par Échantillon

RÉSUMÉ

Les données provenant d'enquêtes contiennent souvent une ou plusieurs observations qui contribuent de façon disproportionnée aux estimés. Ce rapport étudie les différentes méthodes de traitement de telles observations, en insistant surtout sur les problèmes reliés à l'utilisation d'une base de sondage aréolaire. L'enquête nationale sur les fermes est utilisée pour présenter les problèmes et quelques solutions possibles.

1.0 INTRODUCTION

Survey designers take pains to avoid a host of difficulties which are conveniently classified as 'outlier problems': populations are stratified; the largest units are sampled with certainty; sometimes, more than one frame is used. Despite these efforts, 'outliers' still manage to appear in many surveys. The reasons for this may be summarized by the term 'frame inadequacy', which means that either the frame was not very good to begin with or there has been enough change in the population that the design based on the frame no longer reflects, or represents, the current structure of the population. For example, if strata are defined in terms of a measure of size, then the stratum containing the smallest units in the **frame** will now contain units which have increased substantially in size. Similarly, some moderately sized units will have grown enough to qualify as specified units (i.e., units which would have been sampled with certainty had their true size been known). These two examples hint at the confusion which can arise if terms like 'outlier' are interpreted loosely instead of being defined precisely.

In the first example, the unit is assumed to be small in the frame. The stratum containing such units is usually the one with the smallest sampling fraction. As a result, each sampled unit from this stratum 'represents' a large number of other units. If a sampled unit from this stratum has grown to such an extent that it should no longer be in the stratum, then it may have a disproportionate effect on estimates. Note, however, that the unit's size need not classify it as an outlier in the commonly accepted meaning of the term. That is, the raw value for the unit may not be inordinately large; the problem arises when the raw value is multiplied by its 'raising factor' (i.e., the number of farms it 'represented' at the design stage).<*> Thus, if we are estimating a total, the unit is a large contributor to the estimate. To borrow a term from other branches of statistics, we will refer to such units as **influential observations.**

In the second example, the unit would be called an **outlier** if this is defined as an observation whose value exceeds some number (e.g., the number that was used as a lower limit for the stratum of specified units). Note that the observation need not be influential in this case. For example, its contribution to the estimated total may be modest.

Based on the two examples, we can view the problem as a 'stratum jumper' one. In multiple frame surveys, this can be generalized to a 'frame jumper' problem. For example, the National Farm Survey is based on a list frame containing medium and large non-specified farms, and an area frame consisting of smaller farms. As we would expect from the first example, farms in the area frame may become influential. Possible causes are rapid growth of a formerly small operation or of a new operation, and the transfer (e.g. sale) of a list frame farm to someone in the area frame. In some cases, such farms would have appeared in the list frame had up-to-date information been available.

<*> There is no problem if non-sampled units 'represented' by the sampled one also grew significantly in size. However, there is usually no way of knowing this until new information on all units becomes available.

When there are such units in the population, we are faced with two fairly distinct problems<*>. The first is the **detection** of outliers and influential observations. In this report, it will be assumed that the detection has already been done, and we will focus on the second problem, namely, the **treatment** of outliers and influential observations. In addition, the numerical applications of various method will deal exclusively with the estimation of totals. Finally, except for the part of the report which reviews some of the literature on outliers, we will be more concerned with influential observations than outliers.

It should be noted that if there is a set of units in the population with large values for an item, then it can happen that none of the units appears in the sample. This is undesirable since estimates will then be too low. A related idea was alluded to in the footnote on the previous page. It should also be emphasized that most of the methods considered in this report yield biased estimators of totals. However, we also expect that the bias is offset by a reduction in mean square error.

In the next section, various methods of dealing with outliers are discussed. The methods which have been applied to the National Farm Survey are mentioned in subsection 2.5. After some concluding remarks in section 3.0, we summarize three papers on the treatment of outliers.

2.0 TREATMENT METHODS

The National Farm Survey (NFS) will be used to illustrate several methods for dealing with influential observations. This survey consists of an area frame and a list frame. For both frames, each stratum has two replicates. In the area frame, a sample of enumeration areas (EAs) is selected for each replicate. Within each EA, a sample of area segments is selected, and all farms within a sampled segment are enumerated. Thus the contribution of each farm to the replicate total is

(raising factor to replicate level)*(raising factor to EA level)*
(proportion of farm's land inside segment)*(value of item).

The product of the first two components will be called the **total raising factor (TRF)** and the third component will be called the **land weight (wt).** Since there are two replicates, the contribution to the stratum and higher estimates is TRF*wt*value/2. A more precise and detailed description of the NFS is given by G. Davidson, IASMD, '1983 National Farm Survey: Note on the Sample Design and Estimation Procedures', September 1984.

2.1 True Post-Specification

If a unit has been identified as an outlier or an influential observation, then it can be post-

---

<*> They can probably be combined using a Bayesian approach.

specified, i.e., its raising factor (and land weight, if applicable), is changed to 1. This is equivalent to saying that **the unit represents only itself** in the population. Such an extreme measure will usually be inappropriate, except possibly for a list frame unit which was fairly large to begin with and which has grown to such an extent that it is above the cut-off point for specified farms. It will rarely make sense to post-specify area frame units. In summary, post-specification may sometimes be useful for outliers but not for (non-outlier) influential observations.

Note: In the description of the following methods, it is assumed that lowering the raising factor of a unit in the area frame creates a gap of unrepresented area units. Although we will usually describe only one way to 'fill' this gap, there may in fact be other alternatives.

2.2 Self-Representation Within Replicate

This method was used in the AES and FES, which were the precursors of the NFS, and computer programs which implement the method are available.<*> Although the method was used for both area and list frames, we will only consider the area frame, using the NFS as an example.

Let y denote the value of the item. Then the contribution to the replicate total is TRF*wt*y. To obtain the contribution to the stratum total, we must divide this by 2, since there are two replicates per stratum. If we change TRF and wt to 1 for a unit which has been identified as being too influential, then its contribution to the replicate total is now simply y. Now, since the unit 'represented' TRF*wt units before and only 1 unit after the treatment, then there is a 'gap' consisting of TRF*wt-1 units left unrepresented. Therefore, we 'fill the gap' by adding (TRF*wt-1)δ to y, where δ is a weighted average of y values for all non-influential sampled units in the replicate (or segment or stratum). Thus the new contribution to the replicate total is y+(TRF*wt-1)δ and the contribution to the stratum total is half this quantity. Hence the stratum total is reduced by (TRF*wt-1)(y-δ)/2. Note that this is also a measure of how atypical y is.

One property of this method is that the reduction in the estimate based on the area frame can be quite drastic if it is applied to influential observations.<**> Also, as can be seen in the example for Quebec pig estimates in Table 1, it can happen that δ is 0, making the reduction even larger.

<*> The programs also identify the values which are to be subjected to the treatment using a sigma-gap type rule. The programs and the detection/treatment methods are described in a note by Jean-Louis Tambay.

<**> However, the method was originally meant to be applied to **outliers** identified by a sigma-gap type rule.

2.3 Transfer to List Frame

Suppose that an area frame farm has been identified as being an outlier in some sense or as being too influential. Let y be the farm's value for the field under consideration. We will make the following assumptions.

(1) The y value is too large to qualify the farm as belonging to the area frame. At the survey design stage, if up-to-date information were available, a farm with such a value would have been placed in the list frame. We can think of the farm as being an outlier with respect to the area frame.

(2) The 'gap' created if the farm is removed from the area frame is TRF*wt-#f, where #f is the number of farms represented by the 'outlier' which have also grown considerably with respect to their y values (i.e., they would also be influential if they were in the sample). Of course, #f is unknown. If we set #f=1, then this implies that no other area farms in the stratum have grown to a similar size. A reasonable choice may be to take #f=N/(2n) (see the next assumption).

(3) Let N denote the size of a list stratum and let 2n be the number of sampled farms in the stratum (n per replicate). Since each sampled list farm can be thought of as representing N/(2n) farms, then the area farm, once transferred to the list frame, will also be assumed to represent N/(2n) farms.

Keeping these assumptions in mind, we will change the TRF and wt of the farm by reclassifying it into the appropriate list frame stratum. In effect, we are pretending that for this farm, we are back at the design stage.

Consequence 1: The new population stratum size is N + N/(2n), and the new sample stratum size is 2n+1. Hence, the new list raising factor, divided by 2, is the ratio of the two values. It is easy to show that it reduces to N/(2n), the old list raising factor divided by 2. Consequently, it is not necessary to alter the raising factors in the list frame.

Consequence 2: The farm's contribution to the estimated total is

$$Ny/(2n) + (TRF*wt-\#f)\delta/2$$

$$= \text{"list part"} + \text{"area gap part"},$$

where δ is as defined previously.

This method is a reasonable compromise between doing nothing and the previous two methods. One should resort to decreasing the total raising factor to 1 only if the unit is now large enough to meet the criteria for specified farms.<*> This is not likely to happen very often. Of course, 'doing nothing' is also unacceptable (otherwise this study would not be

---

<*> Note that this can be thought of as a special case of the 'transfer to list frame' approach if we think of the specified units as the list stratum containing the largest units. Then both N/(2n) and #f equal 1.

needed!). A strong argument in favour of this method is that it is natural: we are simply doing what would have been done at the design stage if the unit had been of the appropriate size.

2.4 Quantile Methods

Suppose that we are interested in an item y. Divide the frame into one hundred groups (percentiles) as follows: order the frame according to y. The first group will contain the smallest units, whose sum of y values will equal one percent of the frame total. The second group will account for the next one percent, and so forth. This procedure can also provide limits for the percentiles, as well as the number of units in each percentile.

Now suppose that a unit in the sample has been identified as an outlier or an influential observation. We can then find the limits, say y1 and y2, of the percentile containing the outlier's current y value. We then find the number n of sampled farms with y value between y1 and y2, and replace the outlier's raising factor by N/n, where N is the number of units in the frame with y value between y1 and y2.

The above can be thought of as partial poststratification. As in poststratification, we should be using the **current** value of N (or a good estimate of the current value). Unfortunately, if this is not available, then the method is not very useful. When it was applied to the NFS using values of N from the original frame, the results were poor. In one case, n and N were equal, that is, because of growth, the number of farms in the **sample** was equal to the number of farms in the **frame**.<*> Use of deciles, i.e. ten groups, instead of percentiles resulted in only marginal improvement (n=52,N=53). In another example, N/n=1.57 for the percentile method and N/n=1.87 for deciles. Again, these ratios were considered to be too low.

Clearly, good estimates of the current N are needed if this method is to be considered. If such estimates become available, it will also be interesting to go one step further and try complete poststratification of the sample.

2.5 Application

The last two methods were applied to the 1984 NFS pig data (field 610) where the original estimates were judged to be too high, with Quebec being an extreme case. The results are presented in Table 1. In each province, at most two influential observations were identified by subject matter experts (influence was measured by the percent contribution of a farm's value to the estimated provincial total). If two observations were identified, the top contributor was first treated (e.g. row N.S. 1 in Table 1) and then both were treated (e.g. N.S. 1+2). Two treatments were used:

transfer to list frame with #f=N/(2n)

<*> Needless to say, it can also happen that n>N, resulting in a raising factor less than 1.

and

self-representation within replicate.

These were compared to the original NFS estimates, the root estimates (discussed briefly later in this report) and the published figures.

With the exception of Quebec, which we will discuss below, the two treatments brought the totals closer to the published figures. Also note the closeness of the results of the two treatments. The explanation for this is that the list frame raising factors were usually 3 or 4, compared to 1*wt when the self-representation method was used. Relative to the provincial total, the difference between 4*y and 1*wt*y will not be significant for moderate values of y.

In Quebec, the top contributor accounted for 19% of the provincial total<**>. The appropriate list raising factor for this farm is 4. When either method is applied, the result is a provincial estimate which subject matter economists consider to be too low. Based on the published figure for Quebec, the original estimate is equally unacceptable. Thus we are dealing with a problem which cannot be solved using the methods discussed so far. Note that the root estimate for Quebec is relatively close to the published figure.

Despite the Quebec pigs situation, which is a somewhat extreme case, we conclude that the 'transfer to list' method performs adequately. It seems to improve the estimates (in the sense that it brings them closer to what the subject matter economists expected) and is more justifiable than the other methods considered so far.

2.6 Woodruff's Method

Woodruff's method<*> for dealing with large observations was first used in the Monthly Retail Trade Survey in the United States. It was generalized by P. Peskun for the Ontario Hog Survey in the early 1970's (a report is available).

We can think of our population as consisting of a set of 'normal' units and a domain of 'large observations'. Assume that the domain does not change from year to year (or, more generally, from period to period). Woodruff's method requires that if a unit is classified as a member of the domain of large observations this year, then it will be included in next year's sample, even if it would have been rotated out of the sample otherwise. This increases the representation from the domain of large observations. The net effect will be to decrease the variance of the level estimator, while maintaining unbiasedness.

In the second year, we will have two estimates for the domain of large observations, one based on the units retained from last year's sample and the other based on new units, if any, which belong to the domain. The average of these two estimates will be our domain estimate.

<**> TRF=387, wt=1, y=3887

<*> Woodruff, R. S. (1963). The use of rotating samples in the Census Bureau's monthly surveys, JASA 58:454-467.

Note that if the 'domain' really consists of an influential observation (for example, the Quebec pig farm discussed earlier), then it is likely that no similar farm will appear in next year's survey. Consequently, the 'domain' estimate will be the average of the influential observation's contribution and 0. Thus the influence of the observation is cut in half.

Although the method seems reasonable, there are several technical problems which must be solved if it is to be applied to a survey such as the NFS. Some of these were studied by Peskun, who broadened the applicability of the method originally proposed by Woodruff. Among the remaining problems are the following:

- both Woodruff and Peskun define the domain of large observations as the set of all units whose item value exceeds some number. How do we deal with influential observations (as opposed to outliers)? For example, can we replace their 'domain' by a 'domain of large contributors', and if so, how does this affect the theory of the method?
- a current 'outlier' can only be adjusted if it is known that it was also an 'outlier' last year. In practice this means that units have to be in the common portion of the sample to be eligible for classification in the 'outlier' domain. Can we get around this restriction and deal with problems as they appear, without sacrificing unbiasedness?
- if a unit which has been placed in the outlier domain this year (or the last few years) no longer falls in the domain next year, what do we do?

2.7 Winsorization

Winsorization is a method for reducing the effect of outliers on estimators of the mean. It has been studied in the context of sample surveys by Fuller and others. Some of the findings are presented later in this report<*>.

Although Winsorization has apparently not been used for the estimation of population totals, it can be adapted to this purpose. We can proceed as follows: at some level, say the replicate or stratum level, replace the largest value of TRF*wt*y by the second largest value. This results in a biased estimator, but some of the desirable properties of Winsorized means, such as smaller mean square error, may carry over under certain circumstances. It should be easy to study this idea empirically using readily available data (e.g. NFS data).

2.8 Root Estimation

The method of root estimation is currently being studied in IASMD, and preliminary results look promising (see the last column of Table 1). We expect that a report on this topic will be written in the near future, and so we will not discuss it here. The method was introduced by Jenkins, Ringer and Hartley (1973, JASA 68:414-419), who developed some theory for simple random sampling. Much of the current work involves generalization of the theory to two-stage sampling.

<*> See the summary of Fuller's paper.

1984 NFS Core + MTF Pig Estimator (610)

Effect of Two Treatment Procedures

		TRANSFER-TO-LIST METHOD			SELF-REPRESENTATION METHOD				
Province	Original Estimate	Estimate	Difference	Difference /Original (%)	Estimate	Difference	Difference /Original (%)	Published Figure	Root Estimat
N.S. 1	176,723	166,490	10,233	5.8%	165,440	11,283	6.4%	158,000	171,7
N.S. 1+2	176,723	156,335	20,388	11.5%	154,930	21,973	12.3%	158,000	171,7
QUE. 1	3,925,786	3,184,438	741,348	18.9%	3,176,672	749,114	19.1%	3,405,000	3,343,0
QUE. 1+2	3,925,786	2,999,938	925,848	23.6%	2,990,372	935,414	23.8%	3,405,000	3,343,0
MAN. 1	1,091,777	1,069,573	22,204	2.0%	1,069,169	17,156	2.1%	1,028,000	1,058,0
SASK. 1	648,695	632,064	16,631	2.6%	631,539	17,156	2.1%	625,000	630,0
SASK. 1+2	648,695	619,958	28,737	4.4%	618,885	29,810	4.6%	625,000	630,0

JG/lg

3.0 CONCLUSION

The problem of outliers and influential observations in sample surveys is clearly a difficult one, especially when different units have different weights and/or raising factors attached. What little work has been done requires modification if it is to be used in such situations. As an interim solution for multiple frame surveys such as the National Farm Survey, we suggest that the 'transfer-to-list' method be used for the area frame, and that one of the methods studied by Hidiroglou and Srinath (see below) be used for the list frame. In the longer term, it is hoped that the theory behind approaches such as Woodruff's method or root estimation will be generalized and applied, or that completely new solutions will be formulated.

The next few pages contain summaries of three papers on the treatment of outliers. We emphasize that in all cases, the sampled units in a stratum all have the same raising factor, hence the methods cannot be applied directly to the area frame sample of the NFS. We present these summaries to give the reader an idea of the type of work that has been done. Please note that in the summaries, it is assumed that the reader is familiar with various concepts such as the MSE (mean square error) and the 'usual estimator of the total'.

Ernst, L. R. (1980). Comparison of estimators of the mean which adjust for large observations, Sankhya C, 42:1-16.

Note: This paper extends some of the results contained in Searls (1966, JASA, 1200-1204).

Assumptions

- $X_1,\ldots,X_n$ are independent and identically distributed observations from a nonnegative continuous distribution with finite mean and variance.
- t is a predetermined cut-off value.
- r is a predetermined integer in $\{1,2,\ldots,n-1\}$.

Ernst studies seven estimators of the mean:

(1) if $X_i \geq t$, then replace X_i by t when finding the average.
(2) if $X_i \geq t$, then replace X_i by $W*X_i$, $0 \leq W < 1$.
(3) if $X_i \geq t$, then discard X_i.
(4) continue sampling until n observations are less than t; discard those which are greater than or equal to t.
(5) Winsorized mean: replace the r largest observations by $X_{(n-r)}$.
(6) trimmed mean: discard the r largest observations.
(7) replace the r largest observations by W times their values, $0 \leq W < 1$.

Conclusion

Estimator (1) is "best" in the sense that, for optimal t, it is at least as efficient (in the MSE sense) as any of the other six estimators for any t,W,r.

Ernst illustrates this result using the exponential distribution. For this distribution with mean μ, the optimal value of t depends on μ and n. For example, if n=10, then $t=2.1\mu$; if n=100, then $t=3.53\mu$; if n=1000, then $t=5.32\mu$.

Hidiroglou, M. A. and Srinath, K. P. (1981). Some Estimators of a Population Total From Simple Random Samples Containing Large Units, JASA 76:690-695.

Assumptions

- the population $\{Y_1,...,Y_n\}$ contains T units with $Y_i > Y$, T unknown, Y known.
- SRSWOR of size n, denoted by $y_1,...,y_n$, where $y_i > Y$ for $i<t+1$.

Four estimators of the population total Y are compared:

(1) the usual weight N/n is replaced by 1 for the t large observations, and by (N-t)/(n-t) for all other observations.

(2) decrease the usual estimator Y by an amount proportional to the difference between the average of the t large values and the average of the other n-t values.

(3) the usual weight N/n is replaced by r for the t large observations, and by (N-rt)/(n-rt) for all other observations. The value r is chosen to minimize the MSE (r=1 reduces this to (1)).

(4) the usual weight N/n is replaced by T/t for the t large observations, and by (N-T)/(n-t) for all other observations. This requires knowledge of T.

Remarks:

(a) The estimators are compared using the MSE in two ways:
  (i) conditional on the number t of large units observed;
  (ii) conditional on $t \geq 1$, i.e. on the presence of at least one large value in the sample.

(b) The paper also gives the biases of the estimators.

(c) The optimal value of r in (3) depends on several parameters including T.

Conclusions

The theoretical conclusions depend on whether one uses (i) or (ii) and will not be summarized here. Based on the simulation study, (1) is recommended if the sampling fraction and t are both small. Otherwise, (2) is preferred. For moderate to large numbers of outliers, (4) is the best estimator, but its use requires knowledge of T. Similarly, although (3) is an excellent estimator in all circumstances, its use requires knowledge of several parameters.

Fuller, W. A. (1970). Simple estimators for the mean of skewed populations, Prepared for the U.S. Bureau of the Census.

Fuller suggests using estimates based on a Winsorized mean<*> for populations whose right tail resembles the right tail of a Weibull distribution<**> (p. 24-26). Whether or not Winsorization is appropriate can be tested (see bottom of p. 35 and top of p. 36 for an example of a procedure; a graphical procedure is mentioned on p. 41). In any case, Fuller points out that there is little loss in efficiency when Winsorization is applied if the population is such that the mean would be a good estimator (p. 4).

Most of the paper deals with samples consisting of i.i.d. observations from a Weibull distribution. The short section on finite populations, beginning on p. 38, assumes that the population values are a random sample of size N from some distribution, i.e. a superpopulation approach is used. This section of the paper is concerned only with theoretical results.

Several estimators of the mean are compared using samples from two real populations (p. 41). The main conclusion is that **although estimators of the mean based on replacing the largest observation by the second largest one are biased, they have much smaller MSE's than does the sample mean.** The results also indicate that the gain in applying Winsorization to the top two observations is not large enough to be worthwhile.

Note: Winsorized means are also studied by Ernst (1980).

<*> r^{th} Winsorized mean $= \frac{1}{n}\left[\sum_{j=1}^{n-r} X_{(j)} + r\, X_{(n-r)}\right]$

where $X_{(j)}$ is the j^{th} order statistic.

<**> $p(x|\alpha,\beta) = \frac{1}{\alpha\beta} x^{\frac{1}{\alpha}-1} \ell^{-(x^{1/\alpha})/\beta}, \quad x>0,\ \alpha>0,\ \beta>0;$

this reduces to an exponential distribution when $\alpha=1$.