SIMPLE AND ROBUST ESTIMATORS FOR SAMPLING

Beat Hulliger, Swiss Federal Statistical Office
Statistical Methods Unit, SFSO, CH-2010 Neuchâtel, Switzerland
(Beat.Hulliger@bfs.admin.ch)

Key Words: Outliers, one-step estimators, robust weighting.

1 Introduction

Robust estimators for sampling have emerged over the last years. Design based (e.g. Searls 1966, Hidiroglou and Srinath 1981, Oehlert 1985, Fuller 1991, Rivest 1993, Hulliger 1995), model assisted (e.g. Gwet and Rivest 1992) and model based (e.g. Chambers 1986, Chambers 1997) approaches to outlier robust estimation have been explored. However, few practical applications of robust estimators are reported. The practical application may have been hindered by lacking software and by the computational complexity of most of the estimators. Furthermore robust estimators need tuning constants to be chosen, are basically nonlinear and thus create problems when aggregating subpopulations. Outliers are tied to variables rather than to units. Since in practical applications there may be hundreds of variables the proper application of robust estimators becomes cumbersome. Last but not least robust estimators are difficult to explain to the users of statistical information. A more technical problem is that the classical robust estimators have to be adapted to cope with sampling and calibration weights.

The ideal robust estimator for sampling would fulfill the following objectives:

1. Robustness: The reaction to outliers should be mild.
2. Simple to implement (No iteration, at most one auxiliary variable) and simple to explain.
3. One set of weights for all variables and subpopulations.
4. Low variance.
5. Low bias.

In this article we first show how medians, winsorized means and trimmed means may be adapted to sampling weights. Then we discuss one-step W-estimators, which are approximations to M-estimators. The formulation of a one-step robustified ratio estimator contains as special cases a one-step robustified weighted mean and a one-step robustified Horvitz-Thompson estimator (HT-estimator). These one-step robustified estimators may be expressed by weighted means. In other words the robustness aspect may be expressed as an additional weight. A simple variance estimator for these one-step estimators is given.

A strategy for the application of these estimators, covering the choice of tuning constants, the aggregation of subpopulations and the extension to several variables, is discussed.

2 Estimators

Suppose we have a sample S from a population $U = \{1, \ldots, N\}$. Our variable of interest is $y_i, i \in U$. The characteristic to estimate is the population mean $\bar{y}_U = \sum_U y_i/N$. A weight w_i is attached to each observation in S. The weights reflect the inclusion probabilities of the sample design, non-response corrections and calibrations. We assume that $\sum_{i \in S} w_i = N$.

2.1 Weighted mean

The weighted mean is

$$T_M = \frac{\sum_S w_i y_i}{\sum_S w_i}$$

We suppose that the weights are constructed in such a way that under the sample design und under reasonable hypotheses on the non-response mechanisms T_M is approximately unbiased. If w_i is the inverse of the inclusion probabilities π_i then T_M is a Hajek-estimator. We do not consider the pure Horvitz-Thompson estimator for the population mean because in practice it is seldom used.

2.2 Weighted median

A weighted median is calculated as follows: Order the observations $y_{(1)} \leq \ldots \leq y_{(n)}$. Let $w_{[i]}$

I thank Rudi Peters for his comments, Monique Graf for reviewing the paper and Lorenzo Cascioni and Iba Ndiaye for providing the data of the examples.

be the weight of $y_{(i)}$. The partial sums of the weights of the ordered observations are defined as $k_j = \sum_{i=1}^{j} w_{[i]} / \sum_{i=1}^{n} w_i$. In fact, k_j is the estimate of the distribution function of y at the point $y_{(j)}$.

Find the index j_d with

$$j_d = \min\{j : k_j \geq 0.5\}.$$

The weighted median is

$$T_D = \text{med}(y_i, w_i) = y_{(j_d)}.$$

In fact this is the upper median and it may be improved for the estimation of the median. But since we use the median only as an intermediate step we stick with the upper median for simplicity. Note that the weighted median may not be expressed as a simple weighted mean.

2.3 Winsorization

A weighted version of the winsorized mean is defined as follows. Choose $\alpha \in [0, 0.5)$. Find the indices j_l and j_u with

$$\begin{aligned} j_l &= \min\{j : k_j \geq \alpha\} \\ j_u &= \max(\{j : k_j < 1 - \alpha\}, j_l). \end{aligned}$$

The weighted winsorized mean is

$$\begin{aligned} T_W &= \frac{1}{\sum_S w_i} \left(\sum_{j=j_l}^{j_u} w_{[j]} y_{(j)} + \right. \\ & \left. \sum_{j=1}^{j_l - 1} w_{[j]} y_{(j_l)} + \sum_{j=j_u+1}^{n} w_{[j]} y_{(j_u)} \right). \end{aligned}$$

Winsorized means may not be expressed as weighted means with weights adding to 1.

A very simple form of winsorization was proposed by Rivest (1993) for the use in sampling: Only the largest and smallest observations are winsorized. This simple form has the advantage that the robustification effect tends to 0 for large sample size where bias becomes predominant in the mean squared error.

2.4 Trimmed means

Trimmed means simply set the weight of observations outside j_l and j_u to zero, where j_l and j_u are the same indices as for a winsorized mean. More formally

$$u_{[j]} = \begin{cases} 1 & \text{if } \alpha \leq k_j < 1 - \alpha \\ 0 & \text{otherwise.} \end{cases}$$

The weighted trimmed mean is

$$T_T = \frac{\sum_S w_i u_i y_i}{\sum_S w_i u_i}.$$

Trimmed means may have a larger bias than the corresponding winsorized means. On the other hand, the weighted trimmed mean may be expressed as a weighted mean.

Again a simple version of the trimmed mean just trims the smallest and largest observation. A slightly more subtle variant trims $t = \max(\text{int}(\log_{10} n), 1)$ of the smallest and largest observations. While still t/n tends to zero with increasing n some more allowance for trimming is made in large samples by this choice.

2.5 Univariate one-step estimator

We need a starting value to make a step from it. Let T_0 be a first estimate of the population mean. T_0 could be any of the above estimators (weighted mean, weighted median, weighted winsorized mean, weighted trimmed mean) but if we want robustness we may not use the weighted mean as a starting value. We want to declare outliers those observations which are far away from the initial estimate T_0. We therefore define a residual scale by

$$\hat{\sigma} = \text{med}(|y_i - T_0|, w_i)/0.67.$$

This estimator is weighted, thus estimating the residual scale in the population. It is the median absolute deviation MAD in case of equal weights and if T_0 is the median. Again we use the upper median. The estimator $\hat{\sigma}$ could be replaced by, e.g. the interquartile range. However, we stick to it for simplicity.

Now we choose a tuning constant $c > 0$, e.g. $c = 5$ and define robustness weights

$$u_i = \begin{cases} 1 & \text{if } |y_i - T_0| \leq c\hat{\sigma} \\ c\hat{\sigma}/|y_i - T_0| & \text{if } |y_i - T_0| > c\hat{\sigma} \end{cases}.$$

Finally a univariate one-step estimator is defined as the weighted mean

$$T_{0S}(c) = \frac{\sum_S w_i u_i y_i}{\sum_S w_i u_i}.$$

We denote a one-step estimator with an index S for step. The weights u_i correspond to the Huber-ψ-function for M-estimators:

$$u_i = \frac{\psi(y_i - T_0)}{y_i - T_0},$$

where $\psi(x) = \min(\max(x, -c\hat{\sigma}), c\hat{\sigma})$. We use this simple ψ-function throughout this article. The main reason is its simplicity. But it also seems that redescending ψ-functions downweight extreme observations too much.

The iteratively reweighted least squares algorithm for the calculation of M-estimators feeds back T_{0S} into T_0 and iterates until convergence. The one-step estimator T_{0S} is the first robustification step in this algorithm. Here we don't want to iterate in order to keep the procedure simple. The starting value T_0 is important for the behaviour of T_{0S} and the choice of the tuning constant should be adapted to T_0 as well as to the distribution of y. If the distribution of y is skewed to the right T_D is usually smaller than T_T and the choice $T_0 = T_D$ needs rather large values of c, like $c = 5$ or even $c = 10$, while for $T_0 = T_T$ values like $c = 3$ or $c = 5$ might be better suited.

2.6 One-step robustification of the HT-estimator

Now we apply the idea of one-step robustification to the Horvitz-Thompson estimator (cf. (11) in Hulliger 1995). Suppose a positive measure of size x_i is known before sampling for the whole of the population and is supposed positively correlated with important variables of the survey. Denote by x_{U+} the population total of x_i. For a Horvitz-Thompson strategy the weights w_i are the inverse of inclusion probabilities $1/\pi_i$ with $\pi_i = nx_i/x_{U+}$. Thus $w_i = x_{U+}/(nx_i)$ and, conversely, $x_i = x_{U+}/(nw_i)$. (Note that $\sum_S w_i = N$ cannot be guaranteed here.) The Horvitz-Thompson estimator is

$$T_{HT} = \frac{1}{N}\sum_S w_i y_i.$$

The model which inspires the HT-estimator is $y_i = \beta x_i + E_i$, with expectation of the error $\mathbf{E}E_i = 0$ and variance $\operatorname{Var} E_i = x_i\sigma^2$ (cf. Hulliger 1995). Of course, the HT-estimator may be applied without any reference to this model. However the model assists in the development of the robustification. Thus the residual for the robustification of the HT-estimator proposed by the author (Hulliger 1995, Section 3.4) is

$$r_i(\beta) = \frac{y_i - \beta x_i}{\sqrt{x_i}} = \frac{y_i - \beta x_{U+}/(nw_i)}{\sqrt{x_{U+}/(nw_i)}}.$$

Now we replace x_i by $x_{U+}/(nw_i)$ and β by NT_0/x_{U+}. Thus we use NT_0 as our first estimate of y_{U+}. Here T_0 is again an initial estimate of the population mean. We could use $T_0 = T_{HT}$ but we prefer T_D or T_T to obtain some robustness. Assuming for the moment that we still know x_{U+} we may calculate the empirical residual

$$r_i(T_0) = \frac{y_i - NT_0/(nw_i)}{\sqrt{x_{U+}/(nw_i)}}.$$

For the robustification we need an estimate of the scale of the residuals σ or of $\sigma\sqrt{x_{U+}}$. We use the median of the absolute residuals $\hat{\sigma} = \operatorname{med}(|r_i(T_0)|, w_i)/0.67$. Let c be a tuning constant chosen by the statistician, e.g. $c = 5$. Construct robustness weights

$$u_i = \begin{cases} 1 & \text{if } |r_i| \le c\hat{\sigma} \\ c\hat{\sigma}/|r_i| & \text{otherwise.} \end{cases}$$

As is easily seen u_i does not depend on a common factor in r_i. Therefore we may drop the factor $\sqrt{x_{U+}}$ in the denominator of r_i. Thus, in fact, x_{U+} needs not be known at the moment of applying this estimator.

Finally a one-step robustification of the HT-estimator is

$$T_{HTS} = \frac{1}{N}\frac{\sum_S w_i u_i y_i}{\sum_S u_i/n}.$$

The denominator $\sum_S u_i/n$ is itself interesting. It is the average robustness weight. It equals 1 if all observations have their full weight and drops below 1 according to the degree of downweighting for robustness.

We may replace N by $\sum_S w_i$. This corresponds to passing from the Horvitz-Thompson estimator to the Hajek-estimator. The resulting estimator is

$$T_{HS} = \frac{\sum_S w_i u_i y_i}{\sum_S w_i \sum_S u_i/n}.$$

2.7 One-step ratio estimator

We have already introduced auxiliary information for the inclusion probabilities of the Horvitz-Thompson strategy. Now we suppose an auxiliary variable x_i, $x_i > 0, i \in U$, is positively correlated with our variable of interest and that the population mean $\bar{x}_U$ is known. In this situation the classic estimator for $\bar{y}_U$ is the ratio estimator

$$T_R = \bar{x}_U \frac{\sum_S w_i y_i}{\sum_S w_i x_i}.$$

For constant weights $w_i = N/n$ the ratio estimator is the best linear unbiased estimator under the model $y_i = \beta x_i + E_i$ with $\operatorname{Var} E_i = x_i\sigma^2$.

We construct a one-step ratio estimator as follows. Start with a robust estimate β_0 of the slope, e.g.

$$\beta_0 = \hat{\beta}_{\text{med}} = \text{med}(y_i, w_i)/\text{med}(x_i, w_i)$$

We also might choose the weighted median of the slopes $\text{med}(y_i/x_i, w_i)$ as starting value. However, the weighted median of the slopes corresponds to the least-squares estimator $\sum_S w_i(y_i/x_i)$ which assumes an underlying residual variance $\text{Var}\, E_i = x_i^2\sigma^2$. In practice the weighted median of slopes seems to downweight the observations with large x_i too heavily and thus looses too much efficiency.

We now estimate the standard deviation of the residuals by the median of the absolute standardized residuals. The standardized residual is $r_i = r_i(\beta_0) = (y_i - \beta_0 x_i)/\sqrt{x_i}$. We standardize with $1/\sqrt{x_i}$ to keep in line with the assumed underlying model $\text{Var}\, E_i = \sigma^2 x_i$. The residual scale is estimated by the weighted median of the absolute residuals $\hat{\sigma} = \text{med}(|r_i(\beta_0)|, w_i)$.

We define robustness weights

$$u_i = \begin{cases} 1 & \text{if } |r_i| \leq c\hat{\sigma} \\ c\hat{\sigma}/|r_i| & \text{if } |r_i| > c\hat{\sigma} \end{cases}$$

These weights u_i robustify only against extreme residuals. Extreme values of x_i may still have undue influence on the estimate. In practice x_i is often negatively correlated with w_i because the model $\mathbf{E}y_i \propto x_i$ is already taken into account in the sample design. In that case the influence of large x_i is compensated by small weights w_i. For generalized M-estimation of a ratio in sampling see, e.g. ,Gwet and Rivest 1992.

Finally compute a robustness and sampling weighted estimate of slope

$$\hat{\beta}_{RS} = \frac{\sum_{i \in S} w_i u_i y_i}{\sum_{i \in S} w_i u_i x_i}.$$

The final estimate for $\bar{y}_U$ is

$$T_{RS} = \bar{x}_U\, \hat{\beta}_{RS}.$$

2.8 Special cases, domains

If there is no useful auxiliary information we may set $x_i = 1, i \in U$. Then we get $\hat{\beta}_{\text{med}} = \text{med}(y_i, w_i)/1 = T_D$ and $r_i(\beta_0) = (y_i - T_D)$ and therefore $T_{RS} = T_{DS}$. Thus the one-step ratio estimator reduces to the univariate one-step estimator if there is no auxiliary information to predict y_i.

If $x_i = x_{U+}/(nw_i)$ we almost get back the one-step robustified HT-estimator. In that case $\sum_S w_i u_i x_i = \sum_S w_i u_i x_{U+}/(nw_i) = x_{U+} \sum_S u_i/n$ and

$$T_{RS} = \bar{x}_U \hat{\beta}_{RS} = \frac{1}{N} \frac{\sum_S w_i u_i y_i}{\sum_S u_i/n},$$

which is the same form as for the robustified HT-estimator T_{HTS}. The only difference to T_{HTS} is that the starting value $\hat{\beta}_{\text{med}} = \text{med}(y_i, w_i)(n\, \text{med}(w_i))/x_{U+}$ is slightly different from NT_D/x_{U+} because N is estimated by $n\, \text{med}(w_i)$. Thus the robustness weights u_i of this T_{RS} and of T_{HTS} differ slightly.

Because of its relative generality we may use the one-step ratio estimator for programming purposes. As with univariate one-step estimators T_{RS} is the first step for the calculation of a M-estimator with a Huber-ψ-function.

A set of robustifying weights may be fixed and used for robustification in domains. The corresponding ratio will usually be re-estimated by

$$\hat{\beta}_d = \frac{\sum_{S_d} w_i u_i y_i}{\sum_{S_d} w_i u_i x_i}.$$

In order to use a ratio estimator the domain mean of x, i.e. $\bar{x}_{U_d}$ must be known. A variance estimator may be built on the residuals $e_i = y_i - \hat{\beta}_d x_i$. The form of the variance estimator (cf. Section 3) is the same as for the whole sample with the exception that the sums extend over the sample in the domain only. However, such a variance estimator does not estimate the variability induced by the random size of the domain in the strata.

3 Variance estimation

The above estimator $\hat{\beta}_{RS} = \sum_S w_i u_i y_i / \sum_S w_i u_i x_i$ can be written as a solution to an estimating equation involving a residual $r_i(\beta) = (y - \beta x_i)/\sqrt{x_i}$. We derive a variance estimator for the robustified one-step ratio estimator by using this implicit definition of the estimator (cf. Binder 1983). The estimating equation is

$$\sum_S w_i u_i(r_i(\beta_0)) r_i(\beta) \frac{x_i}{\sqrt{x_i}} = 0.$$

The solution to the equation with $\beta_0 = \beta$ is the M-estimator with the Huber-ψ-function. The factor $x_i/\sqrt{x_i}$ stems from the derivative of the residual $r_i(\beta) = (y_i - \beta x_i)/\sqrt{x_i}$ in the minimisation problem that underlies the M-estimator. It cancels with $\sqrt{x_i}$ in the denominator of the standardized residual. For the estimation of the variance we use the unstandardized error

$$e_i = \frac{y_i - \hat{\beta}_{RS} x_i}{\sqrt{x_i}} \frac{x_i}{\sqrt{x_i}} = y_i - \hat{\beta}_{RS} x_i.$$

We treat the robustness weight $u_i = u_i(r_i(\beta_0))$ as an observed variable, i.e. we neglect that it depends on the estimator β_0. We linearise the above estimating equation around $\hat{\beta}_{RS}$ and obtain the variance approximation

$$\mathrm{Var}(\hat{\beta}_{RS}) \approx \frac{1}{(\sum_S w_i u_i x_i)^2} \mathrm{Var}(\sum_S w_i u_i e_i).$$

To estimate $\mathrm{Var}(\sum_S w_i u_i e_i)$ we may use the so-called "With-Replacement" formula and arrive at the following variance estimator for $T_{RS} = \bar{x}_U \hat{\beta}_{RS}$

$$v(T_{RS}) = \frac{(\bar{x}_U)^2 n}{(\sum_S w_i u_i x_i)^2} d^2, \qquad (1)$$

where $d^2 = (\sum_S (w_i u_i e_i - \sum_S w_i u_i e_i / n)^2)/(n-1)$. According to the estimating equation $\sum_S w_i u_i e_i = 0$, but we include the mean of the residuals to have a general formula which applies also for combined ratio estimators. We might include a finite population correction $(1 - n/N)$ with the known risk of underestimating the variance. For other variance functions of the residuals, e.g. $\mathrm{Var}\, E_i = x_i^3 \sigma^2$ the formulae may become more complicated.

For the one-step robustified HT-estimator the unstandardized residual is $e_i = y_i - N T_{HTS}/(n w_i)$. The variance estimator becomes

$$v(T_{HTS}) = \frac{1}{N^2} \frac{n}{n-1} \frac{\sum_S (w_i u_i e_i)^2}{(\sum_S u_i / n)^2}$$

If we know the joint inclusion probabilities we may use the Yates-Grundy-Sen or the Horvitz-Thompson variance estimator instead of the "With-Replacement" formula.

For the one-step robustified weighted mean we get

$$v(T_{MS}) = \frac{n}{n-1} \frac{\sum_S (w_i u_i e_i)^2}{(\sum_S w_i u_i)^2},$$

where $e_i = y_i - T_{0S}$.

In stratified sampling with H strata we might robustify the stratum mean or ratio estimator for each stratum separately. Then the variance of the corresponding one-step robustified estimator may be estimated per stratum and combined to an overall variance estimator in the usual way. Since the biases due to the separate robustification may accumulate across the strata we prefer to robustify the stratified mean for all strata together or the combined ratio estimator. Only if the model of a common ratio for all strata is clearly not adequate we would consider the separate ratio estimator. As a variance estimator for the one-step robustified stratified mean we propose

$$v(T_{MS}) = \frac{1}{(\sum_S w_i u_i)^2} \sum_{h=1}^{H} n_h d_h^2,$$

where $d_h^2 = \sum_{S_h} (w_i u_i e_i - \sum_{S_h} w_i u_i e_i / n_h)^2 / (n_h - 1)$. The variance of the one-step robustified combined ratio estimator may be estimated by

$$v(T_{CRS}) = \frac{\bar{x}_U^2}{(\sum_S w_i u_i x_i)^2} \sum_{h=1}^{H} n_h d_h^2,$$

where d_h^2 is as above, but with the residual corresponding to T_{CRS}.

4 Reference for robustification

Two problems arise when one applies the above estimators to a sample. We have to choose a partition of the population or sample on which we want to robustify. And we have to choose the variable or the set of variables which should be considered.

4.1 Subpopulation level

An outlier may be masked by other extreme observations in the population and appear as an outlier only in a certain subpopulation. For example, a large retail trader may look quite innocent in the whole population. But when looking at retail trade alone it may clearly stick out as an outlier. This phenomenon may happen at any level.

On the other hand an observation may appear innocent in a subpopulation because the variability is large and the subpopulation may have a rather flat distribution. But looking at the whole population the variance may be much lower and the bulk of the data may be much more concentrated. Therefore an observation which looks innocent in a subpopulation may appear as an outlier in the whole population.

If we can decide for a reasonable level of subpopulations that form a partition of the population (a poststratification) then we may apply the robustification at that level and derive results for aggregates by summation. The level of subpopulations should allow for a large enough subsamples to estimate variances well. For example we may use economic activity in sectors defined by a classification of economic activity (like NACE at two digits). We should check whether global outliers are hidden in the subpopulations. A relatively large variance estimate for a subpopulation may give a hint.

Maybe even more important for the choice of the partition is whether there are models that fit the partitions well. The problem is that outlier-robustification is rather model-dependent than model-assisted. It is not possible to speak of outliers without a model for the non-outlying observations and if the model does not fit well at least the majority of the data the loss of efficiency due to the robustification is large.

The advantage of fixing a level of robustification is, that there is one weight for all units in the sample. Totals for aggregates of subpopulations may be calculated by summation. However, if we analyse subpopulations below the chosen robustification level we may encounter problems with "new" outliers. One may argue that these "new" outliers in domains are representative for the whole sample and should not be downweighted. A careful analysis may still suggest that a new robustification on the domain level is necessary. This may lead to different robustness weights and thus to domain estimates which do not aggregate to the corresponding estimates on higher levels of the population.

4.2 Variables

The outlier problem is mainly tied to variables and not to units. The robustness weights u_i depend on the variable y_i considered. Since most surveys have multiple variables the outlier problem is genuinely multivariate. However, it seems rather difficult to convince practitioners to use multivariate estimation methods or to use a specific weight for each variable. How can we proceed without going directly to multivariate robustification?

A first proposal is to derive a set of robustification weights $u_i^{(1)}, \dots, u_i^{(k)}$ for the set of variables $y_i^{(1)}, \dots, y_i^{(k)}$ under consideration. In practice one will often be able to concentrate on a few key variables such that the number of weights to calculate is low. Our proposal is then to use the minimum of the weights per unit as the final weight:

$$u_i^R = \min(u_i^{(1)}, \dots, u_i^{(k)}).$$

This ensures robustness. However we must check with the average robustness weight $l = \sum_S u_i/n$ or alternatively with the weighted average robustness weight $l = \sum_S w_i u_i$ how much weight is lost. Often outliers in one variable turn out to be outliers in other variables, too. Therefore u_i^R may not have a much smaller average robustness weight than any of the $u_i^{(k)}$. If the average robustness weight of u_i^R is too small no simple solution exists and we will have to use multivariate robustification. Of course the variance of the final estimator should be used as a guideline, too. However, the bias which should be considered as a counterweight to variance, is basically unknown.

Once the weight is chosen on the basis of a set of key variables we have to test whether the robustness is sufficient for other variables. Strictly speaking there is no guarantee at all.

5 Validation

In practice the big problem is to judge whether a robust estimator is actually less biased than a non-robust but approximately unbiased estimator. One aspect of the question is that so-called non-representative outliers (cf. Chambers 1986) should not be included in the estimand. Since we seldom know for sure, whether an outlier at hand is representative or not, the robustification may in principle as well lessen or enlarge a possible bias.

The results will always go through a final validation of the results by subject matter statisticians who know the field of application very well. They may compare the results over time and with other statistical and non-statistical informations. From a methodology point of view this external validation is not totally satisfactory but nevertheless it is very important.

There is another way to judge the validity of the approach, in particular whether the choice of the tuning constant c in the esimators is reasonable. Our proposal is to submit the observations with robustness weights $u_i < 1$ to the editors of the survey. They may classify these observations as "no outlier", "possible outlier", "clear outlier". In fact this judgement could be routinely given in the edit and imputation phase. But often the focus of edit and imputation is more on consistency than on outlyingness. To reduce the burden of the editors for the validation one could submit only a subsample of the outliers to their judgement. The aim of the procedure is not to "correct" outliers but simply to tune the robustification. It has the further advantage to give the practitioners a feeling for the robustification and helps to build trust in the estimators. Note that if all possible and clear outliers were edited out in the edit phase of a survey, there might be no work left for robust estimators. Knowing the high costs and long time-delays due to editing, a more balanced integration of robust estimators and editing may save cost and time.

Of course various values of the tuning constant should be tested. The estimate and its variance could be plotted against the tuning constant and often a region can be defined where the choice of the tuning constant is not too sensitive. Hulliger 1995 proposes an adaptive choice of the tuning constant such that an estimate of the mean squared error is minimized.

6 Sampling weights

The weights w_i may contain outliers, too. Looking at the ratio between the maximal and minimal weight $w_{(n)}/w_{(1)}$ one gets a first impression of the spread of the weights. Further measures may be derived from the Lorenz-Curve of the weights.

The inclusion probabilities of a sample design can and should be checked for outliers before the sample design is chosen. Thus from the sample design no outlying weights should emerge. But non-response corrections and calibrations may lead to quite extreme weights and good methods to ensure that no outlying weights are introduced by these corrections still have to be developed. In the mean time a simple winsorization of extreme weights is often used in practice (Hulliger et al. 1997).

In practice the choice of calibration constraints and methods to correct for non-response often take into account preliminary estimations which involve variables that possibly contain outliers. Thus outliers may influence the sampling weights. One has to start with a first choice for the sampling weights and develop robustness weights. Then one should check whether the sampling weights combined with the robustness weights are sufficiently close to calibration constraints and yield results which do not throw doubts on the sampling weights again.

7 Examples

7.1 Simulation with housing rent data

Data on four room appartments from the 1990 census of Switzerland was used to check the estimators with simulations. The sample design is stratified according to economic age of the appartment and uses simple random sampling inside strata. The sample design mimicks the one of the quarterly survey on housing rents and uses the net sample size $n = 1662$ of this survey. The differences in net sampling fractions were mainly due to nonresponse. Therefore the weights are not too different (cf. Table 1). In addition to the whole population of four room appartments a domain of four room appartments of somewhat higher standing was considered.

The characteristic of interest is the population mean of net monthly housing rent: CHF 945.08 for the whole of the population, CHF 965.41 for the domain. The population contains several observations which could be considered outliers, but we include them fully in the estimand. An auxiliary variable, surface, is used in the simulations to construct ratio estimators.

The estimators used are the stratified mean (sme), a univariate one-step estimator from the log-trimmed mean (ts) and a one-step ratio estimator (rs). For both robust estimators the tuning constant was set to $c = 3$. The three estimators were used for the domain mean, too (smed,tsd,rsd). For the robust estimators the robustness weights produced by the estimator on the level of the population were used at the domain mean, too. The variance estimator (1) was used throughout, though clearly it is not the best for the stratified mean. The results for a simulation of 800 random samples are shown in Table 2.

The robust estimators are more efficient in variance than the stratified mean as well for the whole population as for the domain. However, due to their bias, the robust estimators have larger root mean squared error than the stratified mean. We compare the Monte Carlo mean of the variance estimator (mc.meanvar) with the Monte Carlo variance of the estimator. The variance estimators for the robust estimators underestimate, namely 9% and 8% for the whole population, 15% and 12% for the domain. The variance estimator for the stratified mean seems to be approximately unbiased for the whole population but underestimates for the domain. The variability of the variance estimator (mc.varvar) is much larger for the stratified mean than for the robust estimators. This is due to very few exceptional samples with outlying variance estimates.

The Monte Carlo mean of the average robustness weight was 99.75% for the TS-estimator and 99.37% for the one-step ratio estimator. The Monte Carlo variation of these values was very low.

To see the effect of the sample size the same set of simulation was run for sample size $n = 554$, i.e. three times smaller and just large enough to avoid problems with empty domains for some strata. The bias of the robust estimators is less than half of the standard deviation for the population mean but still the robust estimators have larger mean squared error than the stratified mean. On the level of the domain, the bias is of the order of 20% of the standard deviation and the mean squared error of the robust estimators is smaller than for the stratified

Table 1: Population and sample sizes for simulations with housing rents

Age	0-5	6-10	11-20	21 +	all
population	28761	34364	72259	263769	399153
sample	167	210	400	885	1662
weight	172	164	181	298	
domain-pop.	4361	3952	9400	46430	64143
expected domain-sample	29.8	27.3	59.8	189.1	306

Table 2: Simulations with housing rents

Sample size $n = 1662$						
	sme	ts	rs	smed	tsd	rsd
mc.mean	945.19	939.37	935.72	967.73	959.5	953.57
mc.var	115.11	85.76	87.28	885.16	756.09	663.17
mc.rmse	10.73	10.88	13.23	29.84	28.13	28.35
mc.meanvar	115.77	82.31	80.69	825.33	663.27	638.75
mc.varvar	61393.55	31.96	22.68	147973.7	10810.35	7926.1
Sample size $n = 554$						
mc.mean	945.29	939.63	936.40	966.77	959.37	953.29
mc.var	293.07	263.75	259.73	2747.82	2407.36	2176.93
mc.rmse	17.12	17.13	18.31	52.44	49.43	48.21
mc.meanvar	295.94	246.9	240.52	2351.1	1975.25	1889.17
mc.varvar	17285.8	853.93	601.7	2568819.01	272031.02	187387.93

Table 3: Domain estimates for one sample

	ts	usd	drs	ursd
T	953.00	952.94	962.87	955.00
SD(T)	26.45	26.44	26.78	25.83

mean. The variance estimator for the univariate one-step estimator again underestimates by 10% (population) and 24% (domain), the variance estimator for the one-step ratio estimator understimates by 2% (population) and 21% (domain). It seems that the asymptotic approximation behind the variance estimator does not hold too well for subsamples of size 106 as is the case for the domain. Again the variance estimator for the stratified mean has much larger variability than for the robust estimators.

Applying the robust estimators on the level of the domain directly for one of the samples, we obtain the results in Table 3. We compare the robustification on the level of the domain (ts and drs) with the estimators that use the population level robustness weights (usd and ursd). The univariate estimates agree closely while the ratio estimates differ by 7.87, some 30% of the standard deviation.

It seems that the weights developed on the level of the whole sample may serve for the robustification of subdomains. Though, of course, in certain cases there are considerable differences.

7.2 Production Survey

The estimators were applied to preliminary data of the Production Survey 1997. One class of economic activity at two digit level of NACE was chosen to illustrate the methods. However, in the practical application finer partitions were used if a subclass was poorly fitted by the common ratio and, as a consequence, downweighted all together. The sample design is stratified according to size (3 strata) with heavy oversampling of the largest enterprises. For this example the stratum fo the largest enterprises was subsampled and some stratum jumpers were lumped into the corresponding size class. Therefore the results of this example cannot be compared with the official figures. The sample sizes n and population sizes N of the 3 strata are shown in Table 4. The weights were derived from sampling weights with some non-response correction added. The values of the weights range from 2.7 to 189.5.

The variables of interest are production (prod), intermediate consumption (icons) and labour force costs (lfcost). The auxiliary variable is number of full time equivalent jobs. The estimators were the stratified mean (sme), the log-trimmed mean

Table 5: Production survey: estimates

		sme	ltrim	ds	ts	cr	rs	rm
	large c	0	0	40	20	0	20	20
prod	T	2009.54	1973.27	1811.35	1907.29	2009.54	1984.19	1982.67
prod	SD(T)	97.45	88.39	75.67	80.72	62.33	57.32	57.18
prod	$\bar{u}_S$	0	99.03	96.4	98.8	0	99.56	99.52
	moderate c	0	0	20	10	0	10	10
prod	T	2009.54	1973.27	1663.85	1805.33	2009.54	1953.87	1946.00
prod	SD(T)	97.45	88.39	73.02	75.75	62.33	48.46	47.48
prod	$\bar{u}_S$	0	99.03	89.27	96.22	0	98.2	97.73
icons	T	956.28	932.51	780.37	841.36	956.28	915.16	907.81
icons	SD(T)	56.13	50.84	42.92	43.92	48.03	38.46	37.61
icons	$\bar{u}_S$	0	99.03	90.45	96.13	0	98.26	97.79
lfcost	T	898.11	881.94	756.09	817.63	898.11	894.02	894.65
lfcost	SD(T)	43.71	39.19	33.24	34.45	16.63	15.32	15.47
lfcost	$\bar{u}_S$	0	99.03	90.04	96.64	0	99.72	99.79

Table 4: Production survey: sizes

Stratum	small	medium	large	Total
N	19287	2122	1021	22430
n	102	20	291	413

Table 6: Common vs. individual weights

	prod	icons	lfcost
	common weight		
T	1970.84	925.30	892.50
SD(T)	52.25	40.02	15.37
	individual weights		
c	15	15	10
T	1974.11	931.41	894.02
SD(T)	53.53	42.54	15.32

(ltrim), a one-step robustified univariate estimator with the weighted median as starting value (ds) and with starting value ltrim (ts), the combined ratio estimator (cr), a one-step robustified ratio estimator (rs) and a fully iterated M-estimator (rm).

We first tried out some constants to see how much robustification might be suitable (see Table 5). In fact we plotted the function of the difference of the estimator to the stratified mean and the standard deviation of the estimator versus the tuning constant. A tuning constant that yields roughly a difference of half the standard deviation seemed to work quit well. The average robustness weights shown in the tables indicate that for these data rather large tuning constants might be suitable. For the variables prod and icons $c = 15$ might do well. The variable lfcos has somewhat less outliers and $c = 10$ might be more appropriate.

The difference between the one-step ratio estimator and the fully iterated M-estimator is small compared with the difference to the ratio estimator.

We took the minimum of the individual robustness weights u_i for the three variables prod, icons, lfcost from the one-step ratio estimator with tuning constants $15, 15$ and 10 respectively and recalculated the robustness weighted estimators for the variables with this common weight (see Table 6). The average robustness weight of the common weight is 99.74%, very close to the average robust weights for the individual variables. The correlation of the robustness weights $u_i^{(k)}$ is very high. The estimates with the common weight are slightly lower than with individual weights. For prod and icons also the standard deviation is lower than with individual weights.

8 Conclusion

One-step estimators are relatively simple to implement, to explain and to handle. The result of the robustification is an additional robustness weight. One-step ratio estimators give a suitable unified framework for many practical situations including Horvitz-Thompson estimators. External validation and checks by the editors of a survey can help to choose an appropriate tuning constant. For univariate estimators and skew distributions a trimmed mean seems to be better suited as starting value than the weighted median. The choice of an appropriate subpopulation level and the combination of robustness weights for different variables needs careful analysis. In the two examples presented these problems were not too severe. The proposed variance estimators are only slightly more complicated than for the classical estimators. They tend to underestimate the true variance moderately. The application of the robustness weights to domains is possible but again needs careful analysis.

9 Bibliography

Binder, D.A. (1983), On the Variances of Asymptotically Normal Estimators from Complex Surveys, *International Statistical Review*, 51, 279-292.

Chambers, R.L. (1986), "Outlier Robust Finite Population Estimation," *Journal of the American Statistical Association*, 81, 1063-1069.

Chambers, R.L. (1997) Weighting and Calibration in Sample Survey Estimation, *Proceedings of the Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth*, Birkhäuser, Basel.

Fuller, W.A. (1991), *Simple Estimators of the Mean of Skewed Populations*, Statistica Sinica, 1, 137-158.

Gwet, J.-P., Rivest, L.-P. (1992) "Outlier Resistant Alternatives to the Ratio Estimator," *Journal of the American Statistical Association*, 87, 1174-1182.

Hidiroglou, M.A., Srinath, K.P. (1981), "Some Estimators of a Population Total From Simple Random Samples Containing Large Units," *Journal of the American Statistical Association*, 76, 690-695.

Hulliger, B. (1995) Outlier Robust Horvitz-Thompson Estimators, *Survey Methodology*, June 1995, 21/1, 79-87.

Hulliger, B., Ries, A., Comment, T., Bender, A. (1997) Weighting the Swiss Labour Force Survey, *Proceedings of the Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth*, Birkhäuser, Basel.

Oehlert, G.W. (1985), "The Random Average Mode Estimator," *Annals of Statistics*, 13, 1418-1431.

Rivest, L.-P. (1993), "Winsorization of Survey Data," *Proceedings of the ISI 49th Session*, Firenze 1993.

Searls, D.T. (1966), "An Estimator for a Population Mean Which Reduces the Effect of Large Observations," *Journal of the American Statistical Association*, 61, 1200-1204.