Stockholms universitet

Research Report

Department of Statistics

No. 2015:1

Model-Based Value Modifications for Samples from a Skew Population

Olivia Ståhl

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Model-Based Value Modifications for Samples from a Skew Population

Olivia Ståhl

Abstract

Many customary outlier treatment methods, such as for example winsorization, will decrease the mean squared error of a point estimator at the price of a negative bias and lower coverage rates for the corresponding confidence intervals. We propose an alternative, model-based, value modification method that aims at achieving the same decrease in MSE without sacrificing as much in terms of unbiasedness and coverage. The idea is to use a Pareto distribution to model the largest values of the population, as a way to account for extreme values both in and outside the sample. A small simulation study performed on lognormal data shows promising results.

Keywords: Outliers, model-based approach, design-based approach, value modification, winsorization, simple random sampling.

1. Introduction

We address the important problem of estimating the total of a highly skewed finite population variable when the design is simple random sampling without replacement. Design-based methods tend to give reliable results for large samples, as long as the population is fairly symmetric. For highly skewed distributions, model-based procedures are attractive alternatives, at least in theory; empirical results have, however, shown that fully model-based estimators often suffer from too much bias to be practically useful in the finite population context (see for example discussions in Fuller, 1991, and Beaumont and Rivest, 2009). As a way to address this trade-off we propose a class of estimators that combines design- and model-based features to gain the strengths of both approaches.

Lee (1995) gives an introduction to the topic of estimation for skewed finite populations with focus on simple random sampling. (See also Beaumont and Rivest, 2009, for a more general overview.) An important feature of finite population outliers in general is that they are often, in the terminology of Chambers (1986), *representative*. This representativeness is usually interpreted in terms of outliers

detected in the sample representing other equally large values in the population. However, for skewed populations we can also think in terms of representativety of non-sampled units; even when the sample does not contain any extreme values, there may well exist such values in the population. Most outlier treatment methods only account for the first of these two properties. Thus, when faced with large outliers in the sample these are adjusted for using e.g. weight adjustments or winsorization, but potential outliers in the non-sampled part of the population (i.e. non-sampled units that would be qualified as outliers if sampled) are seldom addressed. Two exceptions are the estimators proposed by Fuller (1993) and Balog and Thorburn (2007), which both rely on parametric models for the right tail of the distribution to account for both sampled and non-sampled outliers, and hence for the underlying skewness of the population. Fully model-based approaches (such as for example the one used by Karlberg, 2000) also have this property. In this paper we extend, in particular, the approach of Balog and Thorburn (2007), which means that we will use a Pareto model for the right tail of the distribution but a design based estimator for its bulk.

The Pareto distribution is one of the most commonly used parametric distributions for modeling phenomenon such as income, wealth or turnover; variables which are often highly skewed to the right (see for example Kleiber and Kotz, 2003, for an overview of so called *size distributions*). In particular, it has often been used to model right tail behavior of income distributions, owing to empirical evidence of its appropriateness for this aim; Lydall (1968, page 15) writes: "Experience has shown that in very many cases the upper tail of distributions of total income of individuals or families conforms fairly closely to Pareto's function." Together with its simplicity, this makes us believe that the Pareto distribution is a good candidate for right tail modeling of skewed finite populations in general.

While outlier treatment methods used for sample surveys usually focus at achieving better point estimates, less effort has been given to interval estimation. Chambers and Kokic (1993), in a general overview, even stated that "outlier robust statistical inference about finite population quantities (as opposed to point estimation of these quantities) remains an unsolved problem". In this paper we have a dual focus, attempting to achieve both a small mean squared error for the point estimator and acceptable coverage rates for the corresponding estimated confidence intervals. Our simulation results indicate that although more research is certainly needed on the topic of confidence interval estimation for skewed populations, the approach gives a significant improvement as compared to a specific winsorization strategy.

We confine our account to simple random sampling. However, if outlier domains coincide with strata, the estimators can also be applied directly to stratified simple random sampling designs. Note that even if there is auxiliary data available that allows for stratification, outliers with respect to the study

variable are still a ubiquitous problem within strata. Winsorization is probably the most commonly used method in this situation, and the estimators proposed in this paper thus constitute practical alternatives.

In the next section we introduce notations and some preliminaries, and in Section 3 customary weight- and value-modification methods are reviewed. Section 4 introduces the partly model-based estimators using the Pareto distribution and Section 5 reports from the simulation study.

2. Preliminaries

Following standard convention, we use $y_1, \ldots, y_N$ to denote the finite population values, and $y_1, \ldots, y_n$ for the sample values, where N is the population size and n the size of the sample. Also, $y_{[1]} \leq \ldots \leq y_{[n]}$ denote the sample order statistics. The sampling design is simple random sampling without replacement, and no auxiliary data is assumed to be available.

We will use K to denote the number of large values, or outliers, in the population, and k for the corresponding quantity in the sample. Also, τ denotes a threshold value separating ordinary and large values, so that $K = \sum_{i=1}^{N} I_{[y_i > \tau]}$ and $k = \sum_{i=1}^{n} I_{[y_i > \tau]}$ where $I_{[y_i > \tau]}$ is an indicator variable for whether the value of unit i exceeds τ or not. Further, we distinguish between *absolute* and *relative* outliers; when τ is fixed and k random we refer to the k largest sample values as absolute outliers, and when k is fixed and τ random we refer to them as relative outliers. (Note that in the latter case the outlying values are large as compared to the sample at hand, but not necessarily large compared to the finite population as a whole.)

Our aim is to estimate the population total $T = \sum_{i=1}^{N} y_i$. A design-unbiased point estimator is given by

$$\hat{T}_{HT} = \frac{N}{n} \cdot \sum_{i=1}^{n} y_i \tag{2.1}$$

and a corresponding approximate 95% design-based confidence interval (c.f. Särndal et al., 1992, page 528) can be computed as

$$CI = \hat{T}_{HT} \pm 1.96 \cdot \sqrt{\hat{V}_{HT}}$$

where

$$\hat{V}_{HT} = \left(1 - \frac{n}{N}\right) \frac{s^2}{n} \tag{2.2}$$

and s^2 is the sample variance. The variance estimator in (2.2) is unbiased for the variance of (2.1). (We

use the notation HT for the unbiased point and interval estimators because they are special cases of the Horvitz-Thompson estimators for T and $V(\hat{T}_{HT})$, c.f. Särndal et al., 1992, pages 43-45.)

When the population is skewed to the right, both (2.1) and (2.2) will be highly variable. Consequently, the mean squared error can easily be decreased by abandoning the requirement of unbiasedness. The next section describes common methods of this kind.

3. Weight and value modification methods

For a given non-zero number of sample outliers, $k > 0$, a *reweighted* estimator (e.g. Lee, 1995, page 511) is given by

$$\hat{T}_{RW} = \sum_{i=1}^{n} y_i + (N-n)\left[(1-\omega)\cdot\frac{\sum_{i\leq(n-k)} y_{[i]}}{n-k} + \omega\cdot\frac{\sum_{i>(n-k)} y_{[i]}}{k}\right] \tag{3.1}$$

where $\omega \in [0,1]$ is a weight applied to the sample outliers. Reweighting is usually applied to *absolute* outliers. (But see Ernst, 1980, for a counter-example). If K was known, a post-stratification weight given by $\omega = \frac{K-k}{N-n}$ could be computed, and the unbiased post-stratified estimator would be obtained as a special case of (3.1). Further, if population means and variances below and above τ were known, it would be possible to derive the optimal value of ω with respect to the mean squared error of $\hat{T}_{RW}$ (see Hidiroglou and Srinath, 1981). Without strong auxiliary information, however, more ad-hoc types of approaches, where ω is chosen using e.g. expert judgment, are often used. The weight is then usually restricted to lie within the range $\left[0;\frac{k}{n}\right]$, so that the upper limit recovers the unbiased HT estimator and the lower limit corresponds to implicitly assuming all sample outliers to be unique in the population. (The last estimator is referred to by Gross et al., 1986, as the "surprise estimator", and was first discussed by Kish, 1965.)

Another common technique for outlier adjustment is *winsorization*. A one-sided winsorized estimator appropriate for finite population estimation (Gross et al., 1986; Kokic and Bell, 1994) is given by:

$$\hat{T}_{W} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right)\left[\sum_{i\leq(n-k)} y_{[i]} + k\cdot\tau\right] \tag{3.2}$$

and amounts to replacing the k sample outliers by τ in the "prediction part" of the estimator. A winsorized estimator can be applied to both absolute and relative outliers. When absolute outliers are winsorized, Lee (1995) refers to it as a *censored estimator*, and when relative outliers are adjusted Rivest (1993) termed the method *non-parametric* winsorization.

In the context of infinite populations, Searls (1966) derived an expression for the optimal value of τ for the censored estimator. This value naturally depends on properties of the population (more specifically on the expected values above and below τ, as well as on the tail probability), and can thus only be estimated when strong auxiliary data is available. For non-parametric winsorization, a sample-based replacement value, or "estimate", for τ is needed for (3.2) to be computable. The most common choice is to use the $(n-k)$th sample order statistic, i.e. $\hat{\tau} = y_{[n-k]}$. This amounts to a so called *k-times winsorized* estimator. For infinite populations, Rivest (1994) showed that for a wide range of right skewed distributions, $k = 1$ gives a smaller mean squared error than $k = 2$, for the non-parametric winsorized estimator employing $\hat{\tau} = y_{[n-k]}$. (Larger values of k where not treated by Rivest, but are well known to be less efficient for most populations. See for example Fuller, 1991.) An alternative choice of $\hat{\tau}$ to use in the non-parametric winsorized estimator is:

$$\hat{\tau} = \frac{y_{[n-k]} + y_{[n-k+1]}}{2} \tag{3.3}$$

This estimator is intuitively appealing in our case, because we are interpreting τ as a threshold separating outlying and non-outlying units. This special form of winsorized estimator will be evaluated in the simulation study of Section 5, as a special case of the partly model-based estimators developed next.

4. Design-based estimators which are model-based in the tail

From now on we focus entirely on *relative* outliers, as a simple way to ensure that the modifications we propose are applied to every sample irrespective of whether it includes extreme values or not. We will start by introducing the general forms of the proposed combined point and variance estimators, and then derive four pairs of estimators using properties of the Pareto distribution.

4.1. General forms

A point estimator for T that modifies k relative sample outliers, i.e. the k largest values of the sample,

is given by

$$\hat{T}_{mod} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right) \sum_{i \le (n-k)} y_{[i]} + \left(\frac{N-n}{n}\right) \sum_{i > (n-k)} \tilde{y}_i \tag{4.1}$$

where $\tilde{y}_i$ denotes a value replacing the i:th sample order statistic in the estimator. For example, the non-parametric winsorized estimator is obtained by letting $\tilde{y}_i = \hat{\tau}$ for all i. The reweighted estimator in (3.1) can also be written on the form of (4.1) by using the following replacement values:

$$\tilde{y}_i = \left(\frac{n}{k}\right)\left[\omega \cdot \left(\frac{\sum_{i>(n-k)} y_{[i]}}{k}\right) + (1-\omega) \cdot \left(\frac{\sum_{i \le (n-k)} y_{[i]}}{n-k}\right)\right] - \frac{\sum_{i \le (n-k)} y_{[i]}}{k} \tag{4.2}$$

This implies that the unbiased HT estimator can be recovered from the modified estimator either by setting $\tilde{y}_i = y_{[i]}$ in (4.1), or $\omega = 0$ in (4.2). Further, because the modified point estimator can be re-expressed as

$$\hat{T}_{mod} = \frac{N}{n} \sum_{i=1}^{n} y_{i,mod}$$

where

$$y_{i,mod} = \begin{cases} y_{[i]} & , \text{ for } i \le (n-k) \\ \dfrac{n}{N} \cdot y_{[i]} + \left(1 - \dfrac{n}{N}\right) \cdot \tilde{y}_i & , \text{ for } i > (n-k) \end{cases}$$

a (naive) type of variance estimator is given by

$$\hat{V}_{mod} = \left(1 - \frac{n}{N}\right) \frac{\tilde{s}^2}{n} \tag{4.3}$$

where $\tilde{s}^2$ is the sample variance computed based on the modified values:

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n} \left(y_{i,mod} - \bar{y}_{mod}\right)^2}{(n-1)}$$

Just as (4.1) is a modified version of (2.1), we can think of (4.3) as a modification of (2.2). Hence,

the aim of any value modification procedure will be to produce modified samples that, somewhat loosely speaking, "mimic" the finite population more closely; an adjustment method successful in this respect should result in more efficient estimation of both population total and estimator variance. We will propose one such procedure. First, however, we review properties of the Pareto distribution which will be useful later on.

4.2. Auxiliary truncated Pareto distribution

The probability density function of a type I Pareto distributed random variable is given by (e.g. Forbes et al., 2011):

$$f_y(y) = \frac{\alpha y_{min}^{\alpha}}{y^{(\alpha+1)}} \text{ , for } y \geq y_{min} \tag{4.4}$$

where $\alpha > 0$ and $y_{min} > 0$ are shape and location parameters. A property of the Pareto distribution is *truncation invariance*. This property implies that the conditional density function for y given $y > t$, for a fixed t, is given by:

$$f_{y|y>t}(y) = \frac{\alpha t^{\alpha}}{y^{(\alpha+1)}} \text{ , for } y > t \tag{4.5}$$

From (4.5) one can easily derive the corresponding cumulative distribution function

$$F_{y|y>t}(y) = 1 - \left(\frac{t}{y}\right)^{\alpha} \tag{4.6}$$

and, by solving the equation $F_{y|y>t}(q) = p$ for q, the conditional quantile function:

$$q(p) = \frac{t}{(1-p)^{1/\alpha}} \tag{4.7}$$

This truncated Pareto distribution will later be used as a prediction model for values in the right tail of the population.

Assuming an i.i.d. sample from the unconditional distribution in (4.4), Sveinsson et al. (2002) derived the Maximum Likelihood estimators for α and y_{min} based on only the $r \geq 2$ largest sample order statistics. For α this MLE is given by:

$$\hat{\alpha} = \frac{r}{\sum_{i=n-r+2}^{n}\left[ln(y_{[i]}) - ln(y_{[n-r+1]})\right]} \tag{4.8}$$

By setting $r = n$ in this expression, we obtain the ordinary Maximum Likelihood estimator for α, which supposes an i.i.d. sample of size n from (4.4) and unknown location parameter. In our case, however, knowing that the true population distribution is not perfectly Pareto makes it reasonable to use only the largest sample values to estimate α. In the simulation study we will use $r = k$, but in general r could be any value between 2 and n. (Some further comments on the choice of r are made in the end of Section 5.)

4.3. Four pairs of modified estimators employing the auxiliary model

In this section we will treat the estimated threshold value, $\hat{\tau}$, as fixed, and use the corresponding truncated Pareto distribution to derive different sets of tail predictions. (Note that although $\hat{\tau}$ will have an important role in the prediction formulas, it does not enter into the estimator for $\hat{\alpha}$. Hence, a different choice of $\hat{\tau}$ could be used throughout.) The predictions obtained in this way will then be used as replacement values, $\tilde{y}_i$, in the modified estimators given by (4.1) and (4.3).

The first set of tail predictions to be considered is obtained as k times the expected value of the truncated distribution. For $\alpha > 1$ and $t = \hat{\tau}$, the expectation of y under model (4.5) is given by $\hat{\tau} \cdot \frac{\alpha}{\alpha - 1}$. By estimating α by (4.8), we obtain the corresponding replacement values $\tilde{y}_i = \hat{\tau} \cdot \frac{\hat{\alpha}}{\hat{\alpha} - 1}$, and the explicit form of the modified point estimator for $\hat{\alpha} > 1$ is therefore given by:

$$\hat{T}_{PE} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right)\left[\sum_{i \le (n-k)} y_{[i]} + k \cdot \hat{\tau} \cdot \left(\frac{\hat{\alpha}}{\hat{\alpha} - 1}\right)\right] \tag{4.9}$$

Because the expectation under the truncated model is infinite whenever $\alpha \le 1$, using (4.9), however, involves the risk of obtaining arbitrarily large point and variance estimates. A more robust alternative is obtained from the median of the fitted distribution; by letting $q = \frac{1}{2}$ in equation (4.7), and then replacing α by its estimate, we obtain in this case $\tilde{y}_i = \hat{\tau} \cdot 2^{\frac{1}{\hat{\alpha}}}$. The corresponding point estimator thus has the form:

$$\hat{T}_{PM} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right)\left[\sum_{i \le (n-k)} y_{[i]} + k \cdot \hat{\tau} \cdot 2^{\frac{1}{\hat{\alpha}}}\right] \tag{4.10}$$

Note that, unlike $\hat{T}_{PE}$, $\hat{T}_{PM}$ produces finite predictions even when $\hat{\alpha}$ is smaller than or equal to one.

Modified variance estimators associated with (4.9) and (4.10) can be obtained via the "naive

formula" in (4.3). (Note that if the true tail distribution is approximately Pareto with large α, the variance under the truncated distribution could also be used to derive a variance estimator; however, this model-variance is finite only for $\alpha > 2$ and hence it will not be a practicable approach for type of the highly skewed distributions we are interested in.) Unless the sample values are extremely spread out, the modified sample variance, $\tilde{s}^2$, obtained by using the replacement values of (4.9) or (4.10), is, however, likely to be too small; that is, it will underestimate the true population variance. As a way to achieve better estimates of the variance, we will consider two alternative sets of replacement values with greater within-variability. Inducing additional variability into the modified values is one of four approaches suggested by Little and Rubin (2002, page 75) for obtaining better estimates of the true estimator uncertainty when imputations have been used in the place of true observations. This idea was also used, somewhat implicitly, by both Fuller (1993) and Balog and Thorburn (2007); Fuller used model-based estimates of the largest sample order statistics in one of his estimators, and Balog and Thorburn replaced the largest sample values by quantiles of a fitted Pareto distribution before estimating the finite population parameters. We will consider two approaches that are similar to those employed by Fuller, and Balog and Thorburn, respectively; the first one relies on predictions for the sample order statistics and the other one uses the estimated quantile function.

In the following, we let $\mu_{[1:k]}, \ldots, \mu_{[k:k]}$ denote the expected values of $y_{[n-k+1]}, \ldots, y_{[n]}$ under the truncated model in (4.5), with $t = \hat{\tau}$. By plugging in (4.5) and (4.6) into the general formula for the density of an order statistic (e.g. Casella and Berger, 2002, page 229), and deriving the corresponding expectation (making use of a variable transformation), $\mu_{[j:k]}$, for $j = 1, \ldots, k$, can be readily shown to equal $\mu_{[j:k]} = \hat{\tau} \cdot \frac{\Gamma(k+1)\cdot\Gamma\left(k-j+1-\frac{1}{\alpha}\right)}{\Gamma\left(k+1-\frac{1}{\alpha}\right)\cdot\Gamma(k-j+1)}$ as long as $\alpha > \frac{1}{k-j+1}$. Explicit replacement values can thus be obtained from this expression by estimating α in the same way as above and making an appropriate change of indices; this will give $\tilde{y}_i = \hat{\tau} \cdot \frac{\Gamma(k+1)\cdot\Gamma\left(n+1-i-\frac{1}{\hat{\alpha}}\right)}{\Gamma\left(k+1-\frac{1}{\hat{\alpha}}\right)\cdot\Gamma(n+1-i)}$. The corresponding modified point estimator is, for $\hat{\alpha} > 1$, given by:

$$\hat{T}_{POr} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right) \cdot \left[\sum_{i \le (n-k)} y_{[i]} + \sum_{i > (n-k)} \hat{\tau} \cdot \frac{\Gamma(k+1)\Gamma\left(n+1-i-\frac{1}{\hat{\alpha}}\right)}{\Gamma\left(k+1-\frac{1}{\hat{\alpha}}\right)\Gamma(n+1-i)} \right] \tag{4.11}$$

A nice property of (4.11) is that, as compared to (4.9), it affects *only* the variance estimate; the value of $\hat{T}_{POr}$ will be equal to that of $\hat{T}_{PE}$ but the corresponding variance estimates will be different.

However, this also implies that we again risk to obtain arbitrarily large estimates when $\hat{\alpha}$ is small. Because of this risk, the next approach seems more appropriate for the Pareto case.

To derive the explicit form of the replacement values for "the quantile approach", we start from (4.7) with $p = \frac{j}{k+1}$, for $j = 1, \ldots, k$. As a motivation for this choice of p, note that it corresponds to plugging in the expected values of the order statistics of k independent observations from a uniform distribution on the $[0,1]$ interval. Hence, if the quantile function was known, the corresponding replacement values could be viewed as large sample approximations for $\mu_{[1:k]}, \ldots, \mu_{[k:k]}$; c.f. David and Nagaraja (2003, Chapter 4). These "approximations", unlike the direct estimates, however, have the advantage of being finite for all possible values of α. By estimating α in the same way as before, and again making a change of indices, we obtain replacement values given by $\tilde{y}_i = \hat{\tau} \cdot \left(\frac{k+1}{n+1-i}\right)^{\frac{1}{\hat{\alpha}}}$ and the explicit form of the corresponding modified point estimator is therefore:

$$\hat{T}_{PQ} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right) \cdot \left[\sum_{i \leq (n-k)} y_{[i]} + \sum_{i > (n-k)} \hat{\tau} \cdot \left(\frac{k+1}{n+1-i}\right)^{\frac{1}{\hat{\alpha}}} \right] \tag{4.12}$$

Finally, we note that $\hat{T}_{PM}$ and $\hat{T}_{PQ}$ should be more robust than $\hat{T}_{PE}$ and $\hat{T}_{POr}$ with respect to the tail-model assumption, i.e. they should produce less extreme point and variance estimates. $\hat{T}_{PM}$ is further likely to be even more robust than $\hat{T}_{PQ}$; in particular, for odd k the replacement values implied by (4.10) are all equal to the median of those implied by (4.12). However, although model dependence is often a disadvantage when it comes to estimation of finite population parameters, the whole aim of the present modification procedures is to acknowledge the true skewness of the population and some amount of "dependence" is therefore clearly desirable. Thus, it is not evident which estimator will be best in a practical situation. This trade-off between distributional robustness and efficiency will be explored in the simulation study of the next section. We also note that if α would be assumed to approach infinity, all four sets of estimators proposed in this section reduce to the non-parametric winsorized estimator employing (3.3). (This would correspond to a degenerate truncated Pareto distribution with all probability mass located at the point $\hat{\tau}$, and variance equal to zero.) In a sense, all estimators proposed in this paper can therefore be seen as extensions of a non-parametric winsorization approach.

5. Simulation study

A simulation study on lognormal data was performed to evaluate the performance of the partly model-based estimators for different values of k. The mean of the logarithmed values was set to 0 and their standard deviation, denoted σ, to either 1 or 2. Population size was $10{,}000$. (These values were selected to resemble somewhat realistic situations.) $U = 500$ populations were generated from each parameter specification, and $S = 10{,}000$ simple random samples of sizes $n = 50$, $n = 100$ and $n = 500$, respectively, drawn from each population. (This should be enough to guarantee precision at least for $k \leq 20$ for the estimated bias and mean squared errors reported in Tables 5.2 and 5.3.) To assess the performance of the point estimators, the percent relative root mean square error (RRMSE) and percent relative bias (RB), averaged over the 500 populations, were computed;

$$RRMSE = \left(\frac{100}{U}\right)\sum_{u=1}^{U}\left(\frac{1}{T_u}\cdot\sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{T}_{us}-T_u\right)^2}\right)$$

$$RB = \left(\frac{100}{U}\right)\sum_{u=1}^{U}\left(\frac{1}{T_u}\cdot\left[\frac{1}{S}\sum_{s=1}^{S}\hat{T}_{us}-T_u\right]\right)$$

Here, T_u denotes the finite population total of the u:th population, and $\hat{T}_{us}$ its estimate based on the s:th sample. (Similarily, $\hat{V}_{us}$ will later be used to denote the variance estimate obtained from the s:th sample of the u:th population.)

For all models considered in the simulation study, $\hat{T}_{PE}$ and $\hat{T}_{POr}$ could not be computed for a large proportion of the samples; Table 5.1 includes the average proportions for selected values of k. In the following, performance measures will be reported only when the estimator could be computed for all samples.

Mean squared error results for selected values of k are given in Table 5.2. For the non-parametric winsorized estimator (W), which in light of equation (3.3) can be thought of as winsorizing $(k - 0.5)$ units, the best result is obtained for $k = 2$. (This is consistent with simulation results reported in Rivest and Hurtubise, 1995; they showed that an optimal winsorization strategy for lognormal populations with σ equal to 1 or 2 will result in winsorization of between one and two units.) The PM and PQ estimators perform better than the unbiased estimator for a wide range of k values, and for the smallest k values they also outperform the winsorized estimator with $k = 2$; cases where the MSE is at least as low as for the best winsorized estimator are in bold in Table 5.2. For large sample sizes the modified

estimators perform well also for larger values of k, which is to be expected considering that the number of population values estimated in a model-based way is then smaller given the same value of k.

Table 5.1: Proportion of samples resulting in infinite estimates for PE and POr (in percent). Bold figures indicate zero values.

σ	n / k	2	3	5	10	20	25	30	40
	50	0.21	0.22	0.18	0.19	1.16	4.27	15.21	81.17
1	100	0.11	0.09	0.05	0.01	0.01	0.01	0.02	0.16
	500	3.61	1.42	0.30	0.01	**0.00**	**0.00**	**0.00**	**0.00**
	50	7.14	12.10	21.09	46.49	91.14	98.39	99.87	100.00
2	100	5.42	8.60	13.43	25.68	58.66	75.06	87.51	98.47
	500	2.89	4.03	4.96	5.80	8.08	9.82	11.97	17.88

Comparing the results obtained for different values of σ, we note that the effect of modifying the largest sample values is larger the more skew the distribution. It is also interesting to note that the PQ estimator is very stable with respect to the value of k; the PM estimator is in this particular sense *less* robust than the PQ estimator. The reason is that the PM estimator dampens large estimates more heavily than the PQ estimator, but does not compensate for this by inflating the smaller estimates. Figures 5.1 and 5.2 compare the adjustments made by the PM and PQ estimators by plotting a selection of modified point estimates against the corresponding estimates obtained by the unmodified HT estimator. In Figure 5.1 we have $\sigma = 1$, whereas Figure 5.2 describes the pattern for $\sigma = 2$. For all values of k, large estimates are more heavily down-adjusted by the PM estimator than by the PQ estimator. For k as large as 20, the PQ estimator achieves the desired effect of inflating the smallest estimates. However, for this large value of k some of the largest estimates are also up-adjusted, which probably explains why the mean square error for the PQ estimator is still larger for $k = 20$ than for $k = 5$. The smallest mean square error for the PQ estimator was in most cases obtained for k values in the range $[2; 5]$.

Next, results on the relative bias of the point estimators are reported (Table 5.3). Bold digits again indicate that the corresponding MSE is at least as small as for the best winsorized estimator. We note that the PQ estimator has a smaller bias for most k specifications achieving this low MSE level. In particular, the best PQ estimator (in the sense of achieving the most gains in MSE as compared to the unbiased estimator) always has a smaller bias than the best winsorized estimator (based on $k = 2$). The PM estimator shows a pattern similar to that of PQ, but with a larger bias. As expected, the difference in bias is largest for the more skew distributions. Finally, we note that the average design-bias of the PE and

POr estimators is quite small, indicating that their main problem is that of excessive variance.

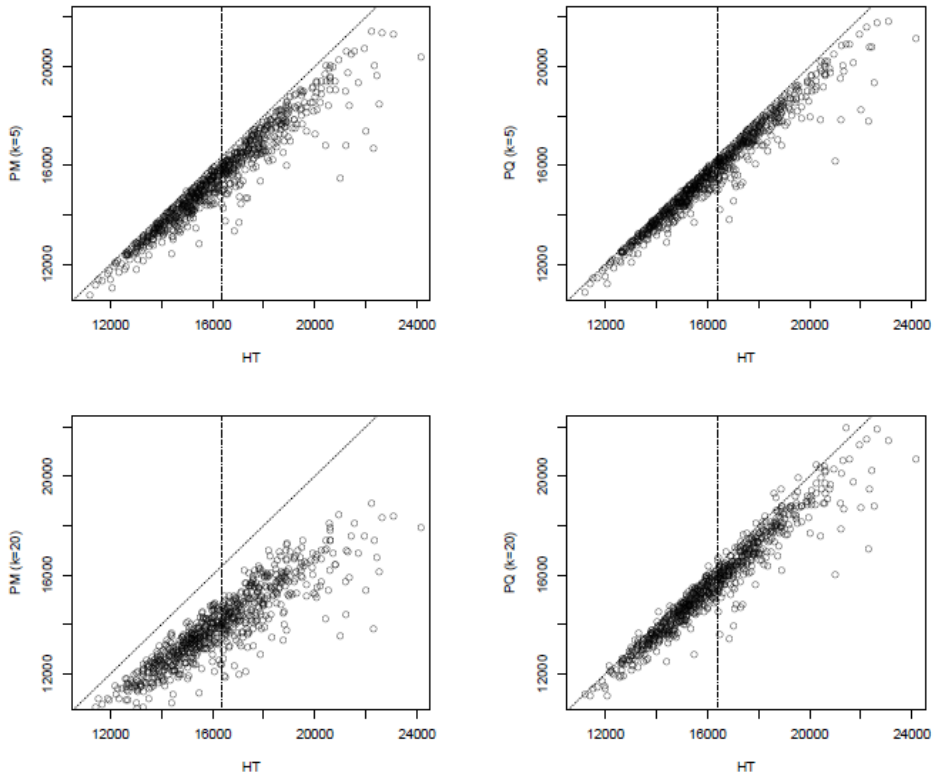

Figure 5.1: Scatter plots of a random selection of point estimates using $k = 5$ and $k = 20$. Log-normal data with $(\mu, \sigma) = (0, 1)$ and sample size $n = 100$. Dotted lines represent zero adjustment (PM=HT or PQ=HT), and average value of true population totals (HT=16488).

Lastly, we consider the interval estimators. Table 5.4 includes average coverage proportions for the estimated confidence intervals obtained by combining (4.1) with (4.3). (Bold digits have the same interpretation as before.) The PQ estimator achieves the highest coverage proportion among the modified estimators, and in particular it outperforms the best winsorized estimator in all cases where the MSE is at an equal level. However, the coverage of the PQ estimator is still lower than for the unbiased estimator. This is because the modified intervals are much shorter than the corresponding unmodified intervals. (This also explains the better coverage properties shown by the POr estimator, where the inherent variation is larger.) Figure 5.3 depicts the average behavior of the PQ and HT intervals. We note

that the best PQ intervals (i.e. those with k in the range 2 to 5) are much shorter than the interval based on the unmodified estimator.

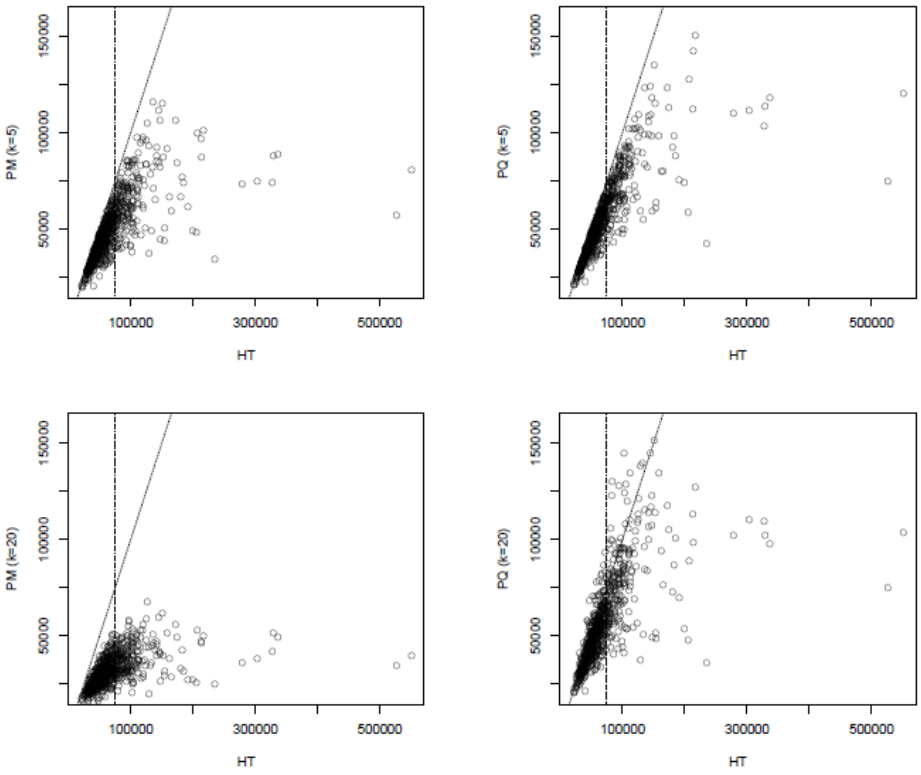

Figure 5.2: Scatter plots of a random selection of point estimates using k = 5 and k = 20. Log-normal data with $(\mu, \sigma) = (0, 2)$ and sample size n = 100. Dotted lines represent zero adjustment (PM=HT or PQ=HT), and average value of true population totals (HT=73874).

Finally, we will comment shortly on the choice of r. Balog and Thorburn (2007) suggested using $r > k$ for estimating α as a way to decrease the variance of the final estimator. For the populations considered in this paper, however, the decreased variability of $\hat{\alpha}$ resulting from using larger values of r did not translate into a smaller variance for the modified estimators. (Note that increasing r generally implies decreasing the value of $\hat{\alpha}$, which has a counteracting variance increasing effect on $\hat{T}_{PM}$ and $\hat{T}_{PQ}$.)

In summary, the simulation results indicate that the modified estimator based on using quantiles

from an auxiliary truncated Pareto model as modified values can achieve both smaller bias and higher coverage, as compared to a non-parametric winsorization strategy with the same level of MSE. However, it should also be noted that while better in an average sense, the maximum error resulting from the partly model-based estimator is likely to be larger than that of winsorization and hence other criteria than MSE and bias should be evaluated in future studies.

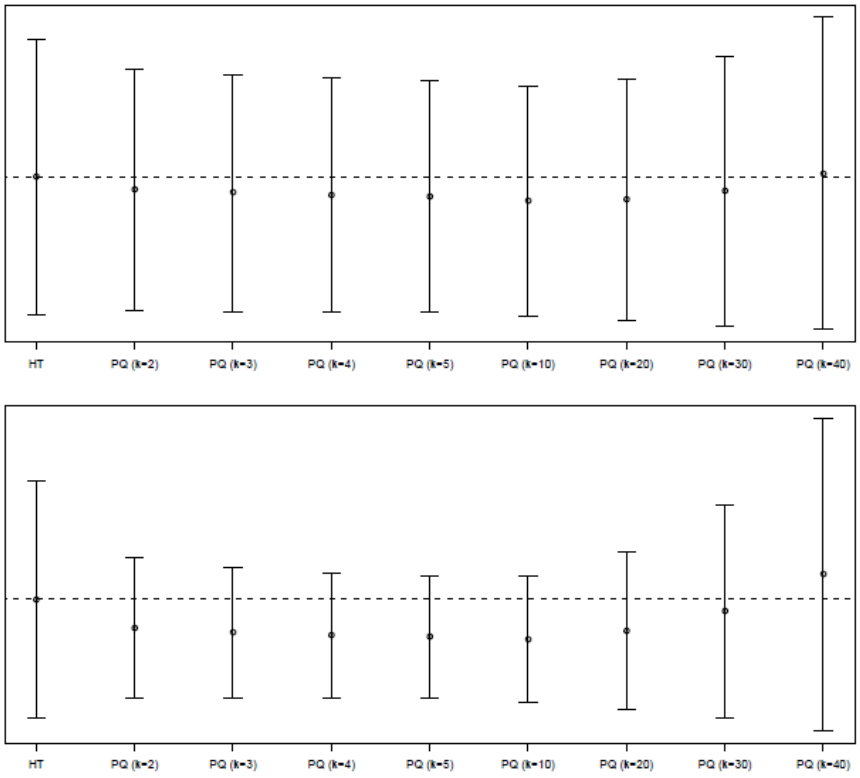

Figure 5.3: Relative length of confidence intervals for the PQ estimators as compared to the unbiased HT estimator, computed as $\overline{\widehat{\mathbf{T}}} \pm 1.96 \cdot \overline{\mathbf{SE}}$ where $\overline{\widehat{\mathbf{T}}}$ is the average over all $\widehat{\mathbf{T}}_{\mathbf{us}}$ and $\overline{\mathbf{SE}}$ the average over all $\sqrt{\widehat{\mathbf{V}}_{\mathbf{us}}}$. Upper panel shows results for lognormal data with $\boldsymbol{\sigma} = \mathbf{1}$ and lower panel for $\boldsymbol{\sigma} = \mathbf{2}$. Sample size $\mathbf{n} = \mathbf{100}$. Dotted lines indicate average value of true population totals.

References

Balog, M, and Thorburn, D. (2007). Extreme value treatment for samples from skew income distributions. *Statistics in Transition*, 8, 1, 139-153.

Beaumont, J.-F, and Rivest, L.-P. (2009). Dealing with Outliers in Survey Data. In D. Pfeffermann, and D. Rao, *Sample Surveys: Design, Methods and Applications, Vol. 29A* (pp. 247-280). Amsterdam: Elsevier North Holland.

Casella, G, and Berger, R. L. (2002). *Statistical Inference.* Pacific Grove, CA: Duxbury.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 396, 1063-1069.

Chambers, R., and Kokic, P. (1993). Outlier robust sample survey inference. *Bulletin de l'Institut international de statistique. Vol. 55, Proceedings of the 49th Session* (pp. 55-72). Firenze: International Statistical Institute.

David, H. A., and Nagaraja, H. N. (2003). *Order Statistics.* Hoboken, N.J.: John Wiley.

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

Forbes, C, Evans, M, Hastings, N, and Peacock, B. (2011). *Statistical Distributions.* Hoboken, N.J: Wiley.

Fuller, W. A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.

Fuller, W. A. (1993). Estimators for long-tailed distributions. *Bulletin de l'Institut international de statistique. Vol. 55, Proceedings of the 49th Session* (pp. 39-53). Firenze: International Statistical Institute.

Gross, W, Bode, G, Taylor, J, and Lloyd-Smith, C. (1986). Some finite population estimators which reduce the contribution of outliers. *Proceedings of the Pacific Statistical Congress* (pp. 386-390). Amsterdam: Elsevier.

Hidiroglou, M. A, and Srinath, K. P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 690-695.

Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. *Metron*, 3-4, 53-80.

Kish, L. (1965). *Survey sampling.* New York: John Wiley and Sons.

Kleiber, C, and Kotz, S. (2003). *Statistical size distributions in economic and actuarial sciences.* Hoboken, NJ: John Wiley and Sons.

Kokic, P, and Bell, P. (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 4, 419-435.

Lee, H. (1995). Outliers in business surveys. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, and P. S. Kott, *Business survey methods* (pp. 503-526). New York: John Wiley and Sons.

Little, R. J, and Rubin, D. B. (2002). *Statistical analysis with missing data (2:nd ed.).* Hoboken, N.J: John Wiley & Sons.

Lydall, H. (1968). *The structure of earnings.* Oxford: Oxford University Press.

Rivest, L.-P. (1993). Winsorization of survey data. *Bulletin de l'Institut international de statistique. Vol. 55, Proceedings of the 49th Session* (pp. 73-89). Firenze: International Statistical Institute.

Rivest, L.-P. (1994). Statistical properties of winsorized means for skewed distributions. *Biometrika*, 81, 2, 373-383.

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

Searls, D. T. (1966). An estimator for a population mean which reduces the effect of large true observations. *Journal of the American Statistical Association*, 61, 316, 1200-1204.

Sveinsson, O. G, Boes, D. C, & Salas, J. D. (2002). Estimation of extreme Pareto quantiles using upper order statistics. *The Extremes of the Extremes: Extraordinary Floods. Proceedings of an international symposium held at Reykjavik* (pp. 289-297). Wallingford: International Association of Hydrological Sciences.

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

Table 5.2: RRMSE of the modified point estimators, relative to RRMSE of the HT estimator.

k	1	2	3	4	5	10	15	20	25	30	40
Lognormal shape parameter $\sigma = 1$, sample size $n = 50$:											
W	0.94	**0.93**	0.97	1.04	1.12	1.58	2.04	2.47	2.88	3.28	4.06
PM	-	**0.92**	**0.92**	**0.93**	0.94	1.09	1.30	1.54	1.80	2.07	2.70
PQ	-	**0.92**	**0.92**	**0.92**	**0.92**	**0.93**	0.95	0.96	1.01	1.19	3.14
Lognormal shape parameter $\sigma = 1$, sample size $n = 100$:											
W	0.96	**0.95**	0.98	1.03	1.09	1.44	1.82	2.19	2.53	2.87	3.49
PM	-	**0.95**	**0.94**	**0.95**	0.96	1.05	1.19	1.35	1.53	1.71	2.08
PQ	-	**0.95**	**0.94**	**0.94**	**0.94**	**0.95**	0.96	0.96	0.97	0.98	1.05
Lognormal shape parameter $\sigma = 1$, sample size $n = 500$:											
W	0.99	**0.98**	0.99	1.01	1.04	1.21	1.42	1.64	1.87	2.09	2.51
PM	-	**0.98**	**0.98**	**0.98**	**0.98**	1.01	1.07	1.13	1.21	1.29	1.47
PQ	-	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	0.99	0.99
PE/POr	-	-	-	-	-	-	1.04	1.05	1.07	1.09	1.14
Lognormal shape parameter $\sigma = 2$, sample size $n = 50$:											
W	0.70	**0.55**	0.58	0.63	0.68	0.84	0.94	1.01	1.06	1.09	1.14
PM	-	**0.54**	**0.53**	**0.54**	0.56	0.68	0.77	0.84	0.91	0.96	1.05
PQ	-	**0.55**	**0.53**	**0.54**	**0.54**	0.56	0.58	0.66	0.91	1.58	10.53
Lognormal shape parameter $\sigma = 2$, sample size $n = 100$:											
W	0.74	**0.62**	0.66	0.71	0.76	0.96	1.09	1.19	1.27	1.33	1.43
PM	-	**0.62**	**0.60**	**0.61**	0.63	0.75	0.85	0.94	1.01	1.08	1.19
PQ	-	**0.62**	**0.61**	**0.61**	**0.61**	**0.62**	0.63	0.65	0.67	0.74	1.07
Lognormal shape parameter $\sigma = 2$, sample size $n = 500$:											
W	0.85	**0.78**	0.81	0.86	0.91	1.16	1.35	1.51	1.65	1.76	1.95
PM	-	**0.78**	**0.76**	**0.77**	**0.78**	0.88	0.99	1.09	1.19	1.27	1.42
PQ	-	**0.78**	**0.77**	**0.77**	**0.77**	**0.78**	**0.78**	**0.78**	**0.78**	0.79	0.79

Table 5.3: Percent relative bias (RB) of point estimators.

k	1	2	3	4	5	10	15	20	25	30	40
Lognormal shape parameter $\sigma = 1$, sample size $n = 50$:											
W	-2.4	**-6.5**	-9.8	-12.7	-15.4	-26.8	-36.3	-44.8	-52.7	-60.2	-74.8
PM	-	**-4.3**	**-5.9**	**-7.4**	-8.8	-15.0	-20.7	-26.2	-31.6	-37.1	-49.3
PQ	-	**-4.0**	**-5.0**	**-5.6**	**-6.1**	**-7.3**	-7.0	-5.5	-2.4	3.6	38.8
Lognormal shape parameter $\sigma = 1$, sample size $n = 100$:											
W	-1.4	**-3.8**	-5.8	-7.6	-9.2	-16.1	-22.0	-27.3	-32.1	-36.6	-45.0
PM	-	**-2.5**	**-3.4**	**-4.3**	-5.1	-8.6	-11.8	-14.8	-17.7	-20.5	-25.9
PQ	-	**-2.3**	**-2.9**	**-3.2**	**-3.5**	**-4.3**	-4.4	-4.1	-3.5	-2.5	0.6
Lognormal shape parameter $\sigma = 1$, sample size $n = 500$:											
W	-0.4	**-1.0**	-1.6	-2.1	-2.6	-4.7	-6.4	-8.1	-9.6	-11.0	-13.6
PM	-	**-0.7**	**-0.9**	**-1.1**	**-1.4**	-2.3	-3.2	-4.0	-4.8	-5.5	-6.9
PQ	-	**-0.6**	**-0.8**	**-0.9**	**-0.9**	**-1.1**	**-1.2**	**-1.2**	**-1.2**	-1.2	-1.0
PE/POr	-	-	-	-	-	-	0.3	0.6	0.8	1.1	1.7
Lognormal shape parameter $\sigma = 2$, sample size $n = 50$:											
W	-13.5	**-32.8**	-42.5	-49.3	-54.5	-70.6	-79.7	-85.6	-89.8	-93.0	-97.0
PM	-	**-25.3**	**-31.9**	**-36.9**	-41.0	-55.1	-64.2	-71.1	-76.6	-81.2	-88.8
PQ	-	**-23.7**	**-27.7**	**-29.9**	**-31.2**	-31.6	-26.3	-15.0	6.4	48.8	460.7
Lognormal shape parameter $\sigma = 2$, sample size $n = 100$:											
W	-9.8	**-24.1**	-31.9	-37.4	-41.9	-56.3	-65.0	-71.3	-76.1	-79.9	-85.6
PM	-	**-18.2**	**-23.2**	**-27.0**	-30.2	-41.5	-49.2	-55.1	-60.0	-64.1	-70.9
PQ	-	**-17.0**	**-20.0**	**-21.7**	**-22.7**	**-23.9**	-22.2	-18.5	-13.2	-5.8	17.7
Lognormal shape parameter $\sigma = 2$, sample size $n = 500$:											
W	-4.1	**-10.5**	-14.4	-17.4	-19.8	-28.5	-34.4	-39.0	-42.7	-45.9	-51.2
PM	-	**-7.6**	**-9.8**	**-11.6**	**-13.2**	-19.0	-23.2	-26.7	-29.6	-32.1	-36.5
PQ	-	**-7.1**	**-8.4**	**-9.2**	**-9.7**	**-10.7**	**-10.7**	**-10.2**	**-9.5**	-8.6	-6.4

Table 5.4: Coverage proportion of interval estimators (in percent).

k	1	2	3	4	5	10	15	20	25	30	40
Lognormal shape parameter $\sigma = 1$, sample size $n = 50$:											
HT	89.3	89.3	89.3	89.3	89.3	89.3	89.3	89.3	89.3	89.3	89.3
W	86.7	**79.3**	71.3	62.7	53.9	16.8	2.4	0.1	0.0	0.0	0.0
PM	-	**83.9**	**81.0**	**77.8**	74.5	54.9	33.5	15.6	5.0	0.9	0.0
PQ	-	**84.5**	**82.9**	**81.8**	**81.0**	**79.4**	80.8	84.2	88.9	93.7	99.4
Lognormal shape parameter $\sigma = 1$, sample size $n = 100$:											
HT	91.4	91.4	91.4	91.4	91.4	91.4	91.4	91.4	91.4	91.4	91.4
W	89.6	**84.8**	79.7	74.2	68.3	37.4	14.4	3.8	0.7	0.1	0.0
PM	-	**87.8**	**86.0**	**84.1**	82.1	70.6	56.8	42.1	28.3	16.9	4.0
PQ	-	**88.2**	**87.2**	**86.5**	**86.0**	**84.8**	85.0	86.0	87.5	89.5	93.7
Lognormal shape parameter $\sigma = 1$, sample size $n = 500$:											
HT	93.9	93.9	93.9	93.9	93.9	93.9	93.9	93.9	93.9	93.9	93.9
W	93.2	**91.5**	89.7	87.9	85.9	74.0	59.6	44.7	31.2	20.2	6.7
PM	-	**92.5**	**91.9**	**91.3**	**90.7**	87.2	83.2	78.5	73.2	67.5	55.2
PE	-	-	-	-	-	-	92.2	91.7	91.2	90.6	89.1
PQ	-	**92.6**	**92.3**	**92.1**	**91.9**	**91.5**	**91.3**	**91.4**	**91.5**	91.7	92.3
POr	-	-	-	-	-	-	94.3	94.8	95.2	95.6	96.3

Table 5.4 Continued

k	1	2	3	4	5	10	15	20	25	30	40
Lognormal shape parameter $\sigma = 2$, sample size $n = 50$:											
HT	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8	68.8
W	62.0	**41.7**	26.6	16.1	9.2	0.3	0.0	0.0	0.0	0.0	0.0
PM	-	**52.3**	**43.6**	**35.7**	28.6	6.8	0.9	0.1	0.0	0.0	0.0
PQ	-	**56.6**	**60.6**	**62.6**	**62.4**	54.6	47.1	43.5	42.6	42.5	42.5
Lognormal shape parameter $\sigma = 2$, sample size $n = 100$:											
HT	73.9	73.9	73.9	73.9	73.9	73.9	73.9	73.9	73.9	73.9	73.9
W	68.1	**50.9**	36.9	25.9	17.6	1.6	0.1	0.0	0.0	0.0	0.0
PM	-	**60.4**	**53.3**	**46.6**	40.3	16.6	5.1	1.2	0.2	0.0	0.0
PQ	-	**63.9**	**67.3**	**68.8**	**68.1**	**57.5**	45.2	35.0	28.2	24.4	22.3
Lognormal shape parameter $\sigma = 2$, sample size $n = 500$:											
HT	82.9	82.9	82.9	82.9	82.9	82.9	82.9	82.9	82.9	82.9	82.9
W	79.3	**68.9**	59.5	50.8	42.8	15.4	4.4	1.1	0.2	0.1	0.0
PM		**75.0**	**71.0**	**67.0**	**63.2**	45.2	30.2	18.9	11.1	6.2	1.7
PQ		**76.9**	**79.0**	**79.3**	**78.2**	**69.7**	**61.5**	**53.8**	**46.5**	39.5	27.0