Partly model-based point estimation for highly skewed populations

Olivia Ståhl

Abstract

We evaluate the merits of estimators for right-skewed data which are motivated by a distributional assumption for the tail of the population. Our results indicate that making use of parametric models to derive the form of estimators can be a fruitful approach for a wide range of right-skewed populations when auxiliary data is not available, especially for small samples.

Keywords: Skewed population, outliers, model-based approach, value modification, simple random sampling, winsorization.

1 Introduction

Although skew distributions are common in business surveys estimators that take this skewness into account in an explicit way, i.e. by modeling the population, are rarely used. The reason is that such estimators risk introducing too much bias. Instead, methods that aim at lowering the mean squared error by damping the effect of large sample "outliers" are common. These methods will by construction result in estimates with a negative bias. The idea explored in this paper is whether modeling only part of the population can yield the same, or a larger, decrease in mean squared error as the dampening type of methods but without the systematic bias component. We consider in particular two estimators, proposed by Fuller (1993) and Ståhl (2015) (see also Balog and Thorburn, 2007) and compare them empirically to the expansion estimator, diffrent types of winsorization approaches as well as to a model based estimator based on a lognormal distributional assumption for the population. We focus on the most basic scenario where no auxiliary data is available and sampling is performed using simple random sampling. In the next section a general framework is presented, and Section 3 describes the different estimators in terms of this framework. Simulation results are presented in Section 4.

2 Framework

Assume that a random sample, $y_1, ..., y_n$, is drawn from a finite population including N values, and that we want to estimate the population total $T = \sum_{i=1}^{N} y_i$ using the sample values. Consider the following class of estimators:

$$\widehat{T} = \sum_{i=1}^{n} y_i + \left(\frac{N-n}{n}\right)\left[\sum_{i\leq(n-k)} y_{[i]} + \sum_{i>(n-k)} \tilde{y}_i\right] \tag{2.1}$$

where $y_{[1]} \leq ... \leq y_{[n]}$ denotes the sample order statistics, $\tilde{y}_i$ is a value replacing the i:th sample order statistic in the estimator (referred to later as the i:th *replacement value*), and k is a fixed non-negative value set a priori by the sampler. We confine ourselves to estimators where k is a pre-defined integer.

3 Estimators

We will compare 6 different estimators in our simulation study, which will be defined in terms of their respective replacement values. As a reference estimator we use the expansion estimator, which can be obtained from (2.1) by setting $k = 0$. As an example of an estimator of the "dampening type", we include the k times winsorized estimator defined by:

$$\tilde{y}_i^W = y_{[n-k]}\,, \forall i \tag{3.1}$$

This winsorized estimator amounts to simply replacing the k largest sample values by the value just below them.

Further, as a first example of an estimator derived using some form of tail modeling, we consider a slightly modified version of the estimator proposed by Balog and Thorburn (2007) and Ståhl (2015). It utilizes the approximate relationship between quantiles of a distribution and the expected values of the order statistics under that same distribution, and the distribution used for this particular estimator is the Pareto. The estimator will be referred to as the Pareto Quantile (PQ) estimator and corresponds to the following replacement values:

$$\tilde{y}_i^{PQ} = \hat{y}_{min} \cdot \left(\frac{n+1}{n+1-i}\right)^{\frac{1}{\hat{\alpha}}} \text{, for } i = (n-k+1), ..., n \tag{3.2}$$

where $\hat{y}_{min}$ and $\hat{\alpha}$ are ML estimates of the Pareto parameters based on the likelihood function which uses only the k largest sample values.

Next, we consider an estimator proposed by Fuller (1993), extended to include also values of k larger than 2. It is derived using a Weibull assumption for the population. We will refer to this estimator as the Order Statistics (OS) estimator since it is based on the idea of trying to predict the values of the order statistics given the Weibull assumption. The replacement values of the OS estimator are given by:

$$\tilde{y}_i^{OS} = \left(y_{[n-k]}^{\hat{\beta}} + c_{k,i} \cdot \hat{\eta}^{\hat{\beta}}\right)^{\frac{1}{\hat{\beta}}} + \frac{\left(1-\hat{\beta}\right)}{2\hat{\beta}^2} \cdot d_{k,i} \cdot \hat{\eta}^{2\hat{\beta}} \cdot \left(y_{[n-k]}^{\hat{\beta}} + c_{k,i} \cdot \hat{\eta}^{\hat{\beta}}\right)^{\left(\frac{1}{\hat{\beta}}-2\right)}$$
$$\text{, for } i = (n-k+1), ..., n \tag{3.3}$$

where $\hat{\beta}$ and $\hat{\eta}$ are the usual ML estimates of the Weibull parameters, and where $c_{k,i}$ and $d_{k,i}$ denote the constants $c_{k,i} = \sum_{j=1}^{k+i-n} \frac{1}{n+j-i}$ and $d_{k,i} = \sum_{j=1}^{k+i-n} \frac{1}{(n+j-i)^2}$, respectively.

We will also include a related estimator suggested by Fuller (1991), which makes use of the Weibull distribution in an implicit way. It will be referred to as the Preliminary Test (PT) estimator and can be defined in terms of the following replacement values:

$$\tilde{y}_i^{PT} = \begin{cases} y_{[i]} & \text{, if } T_{rk} \leq F_{rk} \\ \left[1 + \frac{F_{rk} \cdot (k+1)}{(r-k)}\right] \cdot y_{[n-k]} - \frac{F_{rk} \cdot r \cdot y_{[n-r]}}{(r-k)} + \sum_{i=n-r+1}^{n-k-1} \frac{F_{rk} \cdot y_{[i]}}{(r-k)} & \text{, if } T_{rk} > F_{rk} \end{cases}$$

where

$$T_{rk} = \left(\frac{r-k}{k}\right) \frac{\sum_{i>(n-k)} \left[(n-i+1)\left(y_{[i]} - y_{[i-1]}\right)\right]}{\sum_{i=(n-r+1)}^{(n-k)} \left[(n-i+1)\left(y_{[i]} - y_{[i-1]}\right)\right]}$$

and r is a tuning parameter. (We will use $r = 18$ in the simulations, following a recommendation made by Rivest, 1993.) F_{rk} denotes the 99.5:th quantile of the F distribution with $2k$ and $2(r-k)$ degrees of freedom. The PT estimator is also a type of winsorized estimator, but in this case winsorization is only performed when the sample shows extensive skewness (as measured by the test statistic T_{rk}, which is shown by Fuller to follow an F distribution under the null hypothesis that $\beta = 0$). Note that for the PT estimator some samples will thus not be affected at all by the modification procedure.

Finally, as an example of a fully model based estimator we will consider the approximately unbiased estimator derived by Thorburn (1991), derived under a lognormal assumption for the population. (See also Karlberg, 2000, for an extension.) It will be referred to as the LogN estimator and can be defined in terms of the following replacement values:

$$\tilde{y}_i^{LogN} = exp\left(\bar{z} + \frac{n-1}{2n} \cdot s_z^2 - \frac{1}{4n} \cdot s_z^4\right) \text{ , } \forall i \tag{3.4}$$

where $\bar{z}$ and s_z^2 are the sample mean and variance based on the logarithmed values; $z_i = log\,(y_i), i = 1, ..., n$.

Table 1: Percent relative root mean squared error. Values greater than or equal to 1000 are represented by an asterix. Estimated simulation margin of error within parenthesis.

	k	Weibull $\beta =$ 0.25	Weibull $\beta =$ 0.50	Lognorm $v =$ 1.5	Lognorm $v =$ 2.0	Lognorm $v =$ 2.5	Gamma $a =$ 0.01	Gamma $a =$ 0.05	Gamma $a =$ 0.25
Exp	0	117 (4)	32 (0)	42 (1)	103 (8)	278 (44)	139 (2)	63 (0)	28 (0)
LogN	n	*	814 (99)	31 (0)	51 (0)	78 (2)	99 (1)	100 (2)	*
W	1	66 (1)	29 (0)	31 (0)	50 (1)	74 (3)	99 (1)	58 (0)	28 (0)
	2	66 (1)	30 (0)	32 (0)	50 (1)	72 (3)	94 (1)	61 (0)	29 (0)
PT	2	63 (1)	30 (0)	31 (0)	46 (1)	67 (3)	98 (1)	61 (0)	28 (0)
	3	62 (1)	30 (0)	30 (0)	45 (1)	66 (3)	96 (1)	62 (0)	28 (0)
	4	62 (1)	30 (0)	30 (0)	44 (1)	67 (3)	97 (1)	64 (0)	28 (0)
	5	64 (1)	30 (0)	29 (0)	44 (1)	68 (3)	98 (1)	67 (0)	28 (0)
OS	1	97 (1)	32 (0)	32 (0)	52 (1)	80 (3)	*	*	37 (0)
	2	94 (1)	32 (0)	30 (0)	47 (1)	69 (3)	*	*	46 (0)
	3	93 (1)	32 (0)	30 (0)	46 (1)	67 (3)	*	*	55 (0)
	4	93 (1)	32 (0)	30 (0)	46 (1)	68 (3)	*	*	63 (0)
	5	93 (1)	33 (0)	30 (0)	47 (1)	68 (3)	*	*	69 (0)
PQ	2	67 (1)	29 (0)	31 (0)	51 (1)	76 (3)	101 (1)	58 (0)	28 (0)
	3	62 (1)	29 (0)	30 (0)	47 (1)	69 (3)	92 (1)	57 (0)	28 (0)
	4	62 (1)	29 (0)	30 (0)	47 (1)	69 (3)	89 (1)	57 (0)	28 (0)
	5	62 (1)	29 (0)	30 (0)	47 (1)	69 (3)	510 (248)	57 (0)	28 (0)

4 Simulation study

We performed simulations to find out whether the partly model based estimators have the ability to produce reliable results for different types of populations. To this aim eight types of artificial populations where used, generated from Weibull ($\eta = 1$), lognormal ($m = 0$) and gamma ($b = 1$) distributions (with parameter specifications as in Forbes et al., 2011). Population size was set to $N = 5000$ and sample size to $n = 50$. For each of the data generating models, Monte Carlo estimates of the percent relative root mean squared error (Table 1) and the percent relative bias (Table 2) where computed. The number of populations and samples generated in the simulations varied between models, and the estimated amount of uncertainty due to simulation, as measured by the simulation margin of error ($1.96 \cdot \sqrt{v}$, where v is an estimate of the simulation variance), is reported within parenthesis in the tables.

The tables include results for selected values of k. In general, none of the estimators where very sensitive to the choice of k. The PT estimator worked best for k between 2 and 5 and the non-parametric winsorized estimator for k equal to 1 or 2, which is why the reported output is restricted to those values. For the OS estimator results are reported for k between 1 and 5, and for the PQ estimator for k between 2 and 5.

For the Weibull models, the OS estimators are naturally working quite well, but in some cases their mean squared error performance is still slightly inferior to that of the PT or PQ estimators. Turning to the Lognor-

Table 2: Percent relative bias. Values greater than or equal to 1000 are represented by an asterix. Estimated simulation margin of error within parenthesis.

		Weibull	Weibull	Lognorm	Lognorm	Lognorm	Gamma	Gamma	Gamma
	k	$\beta =$ 0.25	$\beta =$ 0.50	$v =$ 1.5	$v =$ 2.0	$v =$ 2.5	$a =$ 0.01	$a =$ 0.05	$a =$ 0.25
Exp	0	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
LogN	n	*	323 (4)	-1 (0)	-2 (1)	-6 (2)	-98 (1)	-99 (0)	*
W	1	-38 (1)	-9 (0)	-13 (0)	-28 (1)	-47 (2)	-58 (1)	-23 (0)	-6 (0)
	2	-55 (1)	-16 (0)	-21 (0)	-39 (1)	-60 (2)	-82 (1)	-40 (0)	-12 (0)
PT	2	-36 (1)	-2 (0)	-7 (0)	-22 (1)	-45 (2)	-74 (1)	-18 (0)	0 (0)
	3	-44 (1)	-2 (0)	-8 (0)	-25 (1)	-51 (2)	-87 (1)	-27 (0)	-1 (0)
	4	-49 (1)	-3 (0)	-8 (0)	-27 (1)	-54 (2)	-94 (1)	-35 (0)	-1 (0)
	5	-53 (1)	-3 (0)	-9 (0)	-29 (1)	-56 (2)	-96 (1)	-43 (0)	-1 (0)
OS	1	3 (1)	0 (0)	-8 (0)	-19 (1)	-36 (2)	*	*	14 (0)
	2	4 (1)	0 (0)	-12 (0)	-26 (1)	-45 (2)	*	*	25 (0)
	3	6 (1)	1 (0)	-15 (0)	-30 (1)	-49 (2)	*	*	33 (0)
	4	6 (1)	1 (0)	-17 (0)	-32 (1)	-52 (2)	*	*	40 (0)
	5	7 (1)	1 (0)	-18 (0)	-34 (1)	-54 (2)	*	*	45 (0)
PQ	2	-32 (1)	-7 (0)	-10 (0)	-23 (1)	-41 (2)	-51 (1)	-17 (0)	-4 (0)
	3	-41 (1)	-10 (0)	-14 (0)	-29 (1)	-49 (2)	-69 (1)	-27 (0)	-7 (0)
	4	-45 (1)	-12 (0)	-16 (0)	-32 (1)	-52 (2)	-74 (1)	-32 (0)	-9 (0)
	5	-47 (1)	-13 (0)	-17 (0)	-33 (1)	-53 (2)	-58 (1)	-35 (0)	-10 (0)

mal models, the performance of the OS and PQ estimators are approximately equal. The LogN estimator outperforms the expansion estimator here, but it is interesting to note that it is not necessarily better than the partly model-based estimators. Looking at the results for the gamma models, we note that the most skew gamma model stands out in that it is the only case where the PQ estimator is sensitive to the choice of k. For the most skew model ($a = 0.01$), the fully model based LogN estimator worked slightly better than the expansion estimator, but in most other cases it did much worse. (In 3% of the samples the LogN estimator could also not be computed for this model.) For the two most skew models, the bias of the PQ estimator is smaller than that of PT, but for $a = 0.25$ the pattern is the opposite. The performance of the OS estimator is quite similar to that of the PQ estimator when a is equal to 0.25, but goes really bad when a is equal to 0.05 or 0.01.

In summary, the most robust result was obtained for the PT estimator, followed by the PQ estimator. The PT estimator probably has an advantage in its ability to adapt to the shape of the sample, and it would thus be interesting to compare it to a more flexible version of the PQ estimator.

References

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

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical Distributions*. John Wiley & Sons, forth edition.

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. *Invited Paper, Proceedings of the 49th Session of the International Statistical Institute, Firenze*.

Karlberg, F. (2000). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16(3):229–242.

Rivest, L.-P. (1993). Winsorization of survey data. *Invited Paper, Proceedings of the 49th Session of the International Statistical Institute, Firenze*.

Ståhl, O. (2015). Model-based value modifications for samples from a skew population. Research report No 2015:1, Department of Statistics, Stockholm University.

Thorburn, D. (1991). Modelbased Estimation in Survey Sampling of Lognormal Distribution. In *A spectrum of statistical thought, Essays in statistical theory, economics and population genetics in honor of Johan Fellman*, pages 228–243. Swedish School of Economics and Business administration, Helsinki.