RESAMPLING INFERENCE FOR QUANTILE SHARES

Milorad S. Kovacevic, Wesley Yung and Gurupdesh S. Pandher
Wesley Yung, Statistics Canada, Ottawa K1A 0T6

KEY WORDS: Estimating equation, Hierarchical Monte Carlo, Jackknife, Balanced half-sample

1. INTRODUCTION

The inequality of the distribution of income and the polarization of the population are topics of increased interest to society. There exists a large number of different measures for quantifying income inequality and most of their properties have been thoroughly investigated. However, estimation of the variances of such measures has remained elusive for years. Seldom has there been any attempt to provide information about the sampling variability associated with these measures. Such information is of particular interest when income distributions are compared from region to region or across time. Lack of this information confines the role of these measures to that of descriptive devices rather than tools for formal statistical inference.

This paper presents two of these measures, the celebrated Lorenz Curve Ordinates (LCO) and the quantile shares, (QS) also known as income shares. These measures are nonlinear functions of population values and are estimated by complex statistics whose variances are not expressible by simple formulae nor can they be estimated by traditional variance estimation techniques. We have to rely on approximate variance estimation techniques. The major problem with these statistics is that they depend on quantiles. Therefore, their variances should account for two sources of uncertainty: variability of the quantile and variability of the statistic itself assuming that the quantile is fixed.

In a simulation study based on the Canadian Survey of Consumer Finance (SCF) we investigate empirically the performance of different resampling methods and the estimating equations approach for variance estimation.

Section 2 contains definitions of LCO and QS and their complex sample estimators. In section 3 we present three resampling methods: the delete-one-PSU jackknife, the repeatedly grouped balanced half-sample method and the hierarchical Monte-Carlo method. Also, we include the estimating equation method. The simulation study is described in detail in Section 4. Our findings are summarized in Section 5.

2. LORENZ CURVE ORDINATES AND QUANTILE SHARES

We will assume a stratified multistage design with a large number of strata, L, and few primary sampling units (PSU's), $n_h (\geq 2)$, sampled from each stratum. Let w_{hci} be the weight attached to the i-th ultimate unit (a household in the SCF) in the c-th PSU of the h-th stratum such that the size of the population is estimated as $\hat{N} = \sum_s w_{hci}$. We use $\sum_s = \sum_h \sum_c \sum_i$ to denote summation over all ultimate units in the sample incorporating all stages of sampling.

Although the PSU's are usually sampled without replacement, at the variance estimation stage we will treat the PSU's as if they were sampled with replacement. This will lead to a conservative estimate with a small relative bias whenever the first stage sampling fraction is small.

The distribution of income within the population can be depicted by observing the share of income received by the poorest $P=100p$ percent of the population, $0 \leq p \leq 1$. The Lorenz Curve is a graphical representation of that quantity as a function of p and is estimated as

$$\hat{L}(p) = \frac{\sum_s w_{hci} y_{hci} I\{y_{hci} \leq \hat{\xi}_p\}}{\sum_s w_{hci} y_{hci}} = \frac{\hat{\tau}_p}{\hat{\tau}}$$

where $\hat{\xi}_p = inf\{y_{hci}: \hat{F}(y_{hci}) \geq p\}$ is the p^{th} sample quantile and $\hat{F}(y) = \sum_s w_{hci} I\{y_{hci} \leq y\} / \hat{N}$ is the finite population distribution function estimator.

While $L(p)$ is the income share attributed to the poorest P percent of the population, the quantile (income) share is defined as the percentage of total income shared by the population allocated to any

quantile interval $(\xi_{p_1}, \xi_{p_2}]$. The QS is estimated by

$$\hat{Q}(p_1,p_2) = \frac{\sum_s w_{hci} y_{hci} I\{\hat{\xi}_{p_1} < y_{hci} \le \hat{\xi}_{p_2}\}}{\sum_s w_{hci} y_{hci}} = \frac{\hat{\tau}_{p_2} - \hat{\tau}_{p_1}}{\hat{\tau}}.$$

Note that the quantile interval membership is not known beforehand for the population units. However, the size of the interval, expressed in terms of the number of population units in it, is known. For example, the decile interval contains $0.1N$ units. Therefore a quantile interval can be seen as a post-stratum that cuts across the PSU's and strata. On the other hand, the boundaries of the interval are sample dependent and are known conditionally on the realized sample.

3. VARIANCE ESTIMATION

Pioneering work on estimating the variance of these measures has been done by Beach and Davidson (1983). They proposed a transformation of the weighted observations, tacitly assuming the independence of the observations. Their solution can be seen as an approximation in the complex design situation.

In this section, we will briefly review the variance estimation methods used in our simulation study.

3.1 Delete-one-PSU Jackknifing

The jackknife variance estimator is inconsistent for quantiles (Miller, 1974, Kovar, Rao and Wu, 1988). However, recently, under some weak conditions, Shao (1993), showed that asymptotic variances of L-statistics like Lorenz curve ordinates, quantile shares and Gini coefficient can be consistently estimated by delete-one-PSU jackknifing.

The method can be described as the following:

We assume that the estimate of the unknown finite population parameter θ can be expressed as $\hat{\theta} = \mathcal{L}(\hat{F})$, where $\hat{F}$ is the estimated distribution function. The estimate of the distribution function $\hat{F}_{(gj)}$ obtained after removing the j-th sampled PSU of the g-th stratum $(j = 1,...,n_g,\ g = 1,...,L)$ is

$$\hat{F}_{(gj)}(y) = \frac{\sum_s A_{hci}(g,j)\, w_{hci} I\{y_{hci} \le y\}}{\sum_s A_{hci}(g,j)\, w_{hci}},$$

where $$A_{hci}(g,j) = \begin{cases} 1, & h \neq g, \\ \dfrac{n_g}{n_g - 1}, & h = g,\ c \neq j \\ 0, & h = g,\ c = j \end{cases}$$

Then $\hat{\theta}_{(gj)} = \mathcal{L}(\hat{F}_{(gj)})$, and the resulting 'delete-one PSU' jackknife variance estimator for $\hat{\theta} = \mathcal{L}(\hat{F})$ is

$$v_{J1}(\hat{\theta}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\theta}_{(gj)} - \hat{\theta})^2.$$

If $\hat{\theta}$ is substituted by $\hat{\theta}_{..} = \sum_g \sum_j \hat{\theta}_{(gj)}/n$ another variant of the jackknife variance estimate is obtained. We denote it by $v_{J2}(\hat{\theta})$.

3.2 Repeatedly Grouped Balanced Half-Sample (RGBHS) Method

In the grouped balanced half-sample method (GBHS) of variance estimation, the sampled PSU's in each stratum are randomly divided into two groups (halves) and the balanced repeated replication method is applied to the groups. However, Rao and Shao (1993) showed, for stratified random sampling, that this method is asymptotically incorrect in the sense that the associated t-pivotal $t_G = (\bar{y} - \bar{Y}) / \sqrt{v_G(\bar{y})}$ does not converge in distribution to a standard normal distribution. To overcome this difficulty they proposed independently repeating the grouping T times and then taking the average of the resulting T variance estimates. They showed the asymptotic correctness of such an estimator when $\min n_h \to \infty$ and $T \to \infty$. In a small simulation study they found that the method performs well for T as small as 15 in the case of smooth estimators. For an estimator of the population median, the RGBHS method performed better than both the jackknife and GBHS in the sense that the RGBHS had a smaller relative bias and a smaller CV. Although these results were obtained for stratified random sampling, this estimator is expected to perform well in our stratified multistage framework.

We will apply the RGBHS method to variance estimation of LCO and QS. First, in each stratum h, $(h=1,...,L)$, we group the PSU's at random into two halves, h_1 and h_2, containing $m_{h1} = [n_h/2]$ and

$m_{h2} = n_h - m_{h1}$ units, respectively. Then, the group indicator is set to

$$\delta_h^{(r)} = \begin{cases} 1, & h_1 \in r \\ -1, & h_2 \in r \end{cases}$$

where $r = 1,...,R$ denotes a half-sample. The half-samples are balanced if $\sum_{r=0}^{R} \delta_h^{(r)} = 0$ and $\sum_{r=1}^{R} \delta_h^{(r)} \delta_{h'}^{(r)} = 0$, $(h \neq h')$. A minimal set of balanced half-samples is obtained from a Hadamard matrix of order R $(L+1 \leq R \leq L+4)$.

As in the case of the jackknife we assume that the estimator of the unknown finite population parameter θ can be expressed as $\hat{\theta} = \mathcal{G}(\hat{F})$, where $\hat{F}$ is the estimated distribution function. The estimator of the distribution function, $\hat{F}$, based on the r-th half-sample is

$$\hat{F}^{(r)}(y) = \frac{\sum_s A_{hci}^{(r)}(h_1,h_2)\, w_{hci} I\{y_{hci} \leq y\}}{\sum_s A_{hci}^{(r)}(h_1,h_2)\, w_{hci}}$$

where

$$A_{hci}^{(r)}(h_1,h_2) = \begin{cases} 1 + \dfrac{n_h}{2m_{h_2}} \delta_h^{(r)}, & c \in h_1, \\ 1 - \dfrac{n_h}{2m_{h_2}} \dfrac{m_{h_1}}{m_{h_2}} \delta_h^{(r)}, & c \in h_2. \end{cases}$$

When n_h is an even number then $A_{hci}^{(r)}(h_1,h_2)$ is equal to 2 for all units that are in r and 0 otherwise. For an odd case, both half-samples are weighted in a way that $\sum_c A_{hci}^{(r)}(h_1,h_2) = n_h$. This is similar to Fay's idea, given in section 4 of Dippo et al.(1984), of weighting both half-samples within each stratum.

Then $\hat{\theta}^{(r)} = \mathcal{G}(\hat{F}^{(r)})$, and the resulting GBHS variance estimator of $\hat{\theta} = \mathcal{G}(\hat{F})$, based on the t-th random grouping of units, is

$$v_t^{GI}(\hat{\theta}) = \frac{1}{R} \sum_{r_t=1}^{R} (\hat{\theta}^{(r_t)} - \hat{\theta})^2, \quad t = 1,...,T.$$

By repeating the random grouping of units within each stratum T times, computing $v_t^{GI}(\hat{\theta})$ each time and then averaging over T repetitions we obtain the final RGBHS variance estimator

$$v_{RGI}(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} v_t^{GI}(\hat{\theta}).$$

A variant of the GBHS estimator is obtained by replacing $\hat{\theta}$ by $\hat{\theta}_{t.} = \sum_{r_t} \hat{\theta}^{(r_t)}/R$, and will be denoted by $v_{RGI}(\hat{\theta})$.

3.3 Hierarchical Monte Carlo Method

It is felt that the variance estimates of the statistics $\hat{\theta}$ (like LCO and QS) which involve quantiles should incorporate both types of uncertainty arising from the compound randomness of the quantiles and the statistics themselves.

The variance of the estimate $\hat{\theta}$ can be decomposed into

$$V(\hat{\theta}) = E\,V(\hat{\theta} \mid \xi_m) + V\,E(\hat{\theta} \mid \xi_m)$$

assuming that the vector $\xi_m = (\xi_{p_1},...,\xi_{p_m})'$ has some known distribution.

Given the quantiles, ξ_m, the conditional variances of $\hat{L}(p_k|\xi_m)$ and $\hat{Q}(p_{k_1},p_{k_2}|\xi_m)$ can be estimated by the Taylorized variance estimator

$$\hat{V}(\hat{\theta} \mid \xi_m) = \sum_h \frac{n_h}{n_h - 1} \sum_c (a_{hc} - \bar{a}_h)^2$$

where $a_{hc} = \sum_i w_{hci}\, a_{hci}$ and $\bar{a}_h = \sum_c a_{hc}/n_h$. The a_{hci} variates are defined as

$$a_{hci} = \frac{1}{\hat{\tau}} \left[y_{hci} I\{y_{hci} \leq \xi_{p_k}\} - y_{hci} \hat{L}(p_k) \right]$$

for LCO, and

$$a_{hci} = \frac{1}{\hat{\tau}}\left[y_{hci} I\{\xi_{p_{k_1}} \le y_{hci} \le \xi_{p_{k_2}}\} - y_{hci}\hat{Q}(p_{k_1}, p_{k_2})\right]$$

for QS.

Under certain regularity conditions Francisco and Fuller (1991) extended the Bahadur representation of the sample quantiles to a general survey design and proved the asymptotic multivariate normality of m sample quantiles in the case of a single stage, stratified cluster sampling showing that

$$[\hat{D}\,\hat{\Omega}\,\hat{D}]^{-\frac{1}{2}}(\hat{\xi}_m - \xi_m) \to N_m(\mathbf{0}, I) \qquad (3.1)$$

where $\hat{D} = diag(\hat{d}_1,...,\hat{d}_m)$ and for $1 \le k \le m$,

$$\hat{d}_k = \frac{\hat{\xi}^U_{p_k} - \hat{\xi}^L_{p_k}}{2z_{\alpha/2}\hat{S}_k}$$

with $\hat{\xi}^U_{p_k} = \hat{F}^{-1}(p_k + z_{\alpha/2}\hat{S}_k)$ and $\hat{\xi}^L_{p_k} = \hat{F}^{-1}(p_k - z_{\alpha/2}\hat{S}_k)$. The estimated covariance matrix of $\left(\hat{F}(\hat{\xi}_{p_1}),...,\hat{F}(\hat{\xi}_{p_m})\right)'$ is $\hat{\Omega}$ and $\hat{S}_k$ is the square root of the corresponding diagonal element of $\hat{\Omega}$.

If $\xi_m^{(b)}$ $(b = 1,...,B)$ are replicates drawn independently from the distribution (3.1), the variance of $\hat{\theta}$ can be estimated by the Monte-Carlo estimator

$$\hat{V}_{MC}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} \hat{V}(\hat{\theta}^{(b)} | \xi_m^{(b)}) + \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}.)^2$$

where $\hat{\theta}^{(b)}$ is the value of estimate (either LCO or QS) obtained for the b-th replicate of ξ_m, and

$$\hat{\theta}. = \sum_{b=1}^{B} \hat{\theta}^{(b)}/B.$$

This approach was used by Mantel and Singh (1991) for the estimation of the standard errors for low income proportions, as well as by Hamilton (1986) in the context of a standard error for the estimated state vector in a state space model.

3.4 Estimating Equations Approach to Variance Estimation

The estimating equation (EE) approach of Binder (Binder, 1991, Binder and Kovacevic, 1993, Binder and Patak, 1994), unlike the resampling methods, is not computationally intensive. It provides formulas for asymptotic variance which are easy to program despite their complicated appearance.

Applying the EE methodology as given in Binder and Kovacevic (1993) we obtain expressions for the approximate variance estimators of the LCO and QS as

$$v_{EE} = \sum_h \frac{n_h}{n_h - 1}\sum_c \left(u^*_{hc} - \bar{u}^*_h\right)^2$$

where $u^*_{hc} = \sum_i \tilde{w}_{hci} u_{hci}$, $\bar{u}^*_h = \sum_c u^*_{hc}/n_h$, and $\tilde{w}_{hci}$ is a normalized weight.

For the LCO, u^* variates are defined as

$$u^*_{hci} = \frac{1}{\hat{\mu}}\left[(y_{hci} - \hat{\xi}_p) I\{y_{hci} \le \hat{\xi}_p\} + p\,\hat{\xi}_p - y_{hci}\hat{L}(p)\right]$$

and for QS

$$u^*_{hci} = \frac{1}{\hat{\mu}}\left[(y_{hci} - \hat{\xi}_{p_2}) I\{y_{hci} \le \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1}) I\{y_{hci} \le \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\hat{Q}(p_1, p_2)\right]$$

where $\sum_s w_{hci}$

4. SIMULATION STUDY

In order to study the different methods of variance estimation for two non-smooth statistics, LCO and QS, we use a synthetic population based on the Ontario sample for Canadian Survey of Consumer Finance in 1988. The sample size was 7474 households situated in 525 PSU's from 91 strata. By collapsing some strata, we end up with the micro population of 40 strata with 525 PSU's and 7474 households. For each household we have a nonnegative value of annual income. The true values of the parameters of interest were computed from this population. We generated A=5000 independent stratified single stage cluster samples with total sample size of n = 108 PSU's. Using Neyman allocation, the resulting sample sizes were between 2

and 6 PSU's per stratum. PSU's were selected with probability proportional to size with replacement.

The empirical MSE (EMSE) of the parameters of interest were calculated using 10,000 independent samples using the sample design explained above.

We considered the following quantile shares: *Q(0,0.1)*, *Q(0,0.2)*, *Q(0.2,0.4)*, *Q(0.4,0.6)*, *Q(0.6,0.8)*, *Q(0.8,1)*, *Q(0.9,1)*, *Q(0.95, 1)* and the corresponding Lorenz curve ordinates.

From each sample we computed the estimators and the following variance estimators:

- two jackknife estimators, v_{J1},v_{J2},
- two RGBHS estimators, v_{RG1}, v_{RG2}
- the hierarchical Monte Carlo estimator, v_{MC}
- the estimating equation estimator v_{EE}.

The jackknife variance estimators were based on 108 jackknife replicates. For the RGBHS we have used 41 rows of a 44x44 Hadamard matrix along with T=3 repetitions giving a total of 123 replicates. In the case of the hierarchical Monte Carlo method we used B=100 replicates.

In order to evaluate the performance of the methods we computed from simulated samples the following measures:

- the relative bias to assess the accuracy of estimators;
- the coefficient of variation of the variance estimators to assess their precision
- 2-tailed 95% and 1-tailed 97.5% upper and lower confidence intervals to evaluate the coverage rates of confidence intervals.

5. SUMMARY AND CONCLUSIONS

The results of our simulation study are presented in the following tables. The results for the hierarchical Monte Carlo method showed a large overestimation and were very suspect. Thus, the numbers for the hierarchical Monte Carlo method will not be reported as we are still investigating the cause of this discrepancy. Table I reports the relative bias of our estimators for 3 LCO's and 3 QS's. We report only one jackknife variance estimator and one RGBHS estimator, since the results for the two variants are similar.

From Table I, we see that all three methods, jackknife, RGBHS, and EE, appear to track the EMSE fairly well. It is interesting to note that the jackknife estimates the EMSE's of the Lorenz Curve Ordinates better than that of the Quantile Shares, while the other two methods behave similarly for both income inequality measures.

Table I
Relative Bias of Variance Estimators

	v_{J1}	v_{RG1}	v_{MC}	v_{EE}
L(0.1)	24%	−21%	–	−18%
L(0.4)	8%	−3%	–	−10%
L(0.6)	3%	−15%	–	−15%
Q(0,0.1)	27%	−19%	–	−15%
Q(0.2,0.4)	48%	1%	–	−7%
Q(0.4,0.6)	100%	−18%	–	−15%

Table II presents the CV's of our estimators. The jackknife is slightly more variable than the RGBHS or the EE methods. Again, the jackknife behaves better for the Lorenz Curve Ordinates than for the Quantile Shares, and the RGBHS and EE behave similarly for both inequalities.

Table II
CV's of Variance Estimators

	v_{J1}	v_{RG1}	v_{MC}	v_{EE}
L(0.1)	80%	34%	–	39%
L(0.4)	67%	43%	–	38%
L(0.6)	64%	51%	–	46%
Q(0,0.1)	82%	35%	–	41%
Q(0.2,0.4)	143%	52%	–	43%
Q(0.4,0.6)	230%	59%	–	58%

Table III presents coverage rates of nominal 95% confidence intervals. In general, the three methods perform well, with the coverage for the Lorenz Curve Ordinates slightly better than for the Quantile Shares. The jackknife intervals tend to have higher than nominal coverage rates for the Quantile Shares while the remaining two methods tend to have lower than nominal coverage rates for the Quantile Shares. The error rates

of the one-sided intervals, not presented here, show similar results.

Table III
Coverage Rates of 95%
Confidence Intervals

	v_{JI}	v_{RGI}	v_{MC}	v_{EE}
L(0.1)	96%	93%	–	93%
L(0.4)	95%	94%	–	94%
L(0.6)	95%	94%	–	94%
Q(0,0.1)	96%	92%	–	92%
Q(0.2,0.4)	97%	94%	–	93%
Q(0.4,0.6)	98%	91%	–	91%

In conclusion, we see that for the Lorenz Curve Ordinates, the three methods, jackknife, RGBHS, and EE, perform well with the jackknife performing slightly better in terms of the relative bias and the coverage rates but tending to be more variable. For the Quantile Shares, the RGBHS and EE approaches perform comparably and slightly better than the jackknife in terms of relative bias and variability.

REFERENCES

Beach, C.M. and Davidson, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. ***Review of Economic Studies***, 50, 723-735.

Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. ***Proceedings of the American Statistical Association, Survey Research Methods Section***, 34-42.

Binder, D.A. and Kovacevic, M.S. (1993). Estimating Some measures of Income Inequality from Survey Data: An Application of the Estimating Equation Approach. ***Proceedings of the American Statistical Association, Survey Research Methods Section (to appear).***

Binder, D.A. and Patak, Z. (1994). Use of estimating functions for interval estimation from complex surveys. ***Journal of the American Statistical Association***, 89, 1035-1043.

Dippo, C.S., Fay, R.A and Morgenstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. ***Proceedings of the American Statistical Association, Survey Research Methods Section*** , 489-494.

Francisco, C.A. and Fuller, W.A. (1991). Quantile estimation with a complex survey design. ***Annals of Statistics*** 19, 454-469.

Hamilton, J.D. (1986). A Standard Error for The Estimated State Vector of a State-Space Model. ***Journal of Econometrics*** 33, 387-397

Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. ***The Canadian Journal of Statistics*** 16, 25-45.

Mantel, H and Singh, A.C (1992). Memo on the Standard Error Estimation of the LIL and LICOs. ***Statistics Canada.***

Miller, R.G. (1974). The Jackknife - a review. ***Biometrika*** 61, 1-15

Rao, J.N.K. and Shao, J. (1993). On Balanced Half-Sample Variance Estimation in Stratified Sampling. (preprint)

Shao, J. (1993). Inferences Based on L-statistics in Survey Problems: Lorenz Curve, Gini Family and Poverty Proportion. In ***Proceedings of the Workshop on Statistical Issues in Public Policy Analysis***, Carleton University and University of Ottawa