Published in Journal of Statistical Planning and Inference 141, issue 8, 2976-2985, 2011
which should be used for any reference to this work

Statistical inference for the quintile share ratio

Matti Langel *, Yves Tillé

Institut de Statistique, Université de Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland

A B S T R A C T

In recent years, the Quintile Share Ratio (or QSR) has become a very popular measure of inequality. In 2001, the European Council decided that income inequality in European Union member states should be described using two indicators: the Gini Index and the QSR. The QSR is generally defined as the ratio of the total income earned by the richest 20% of the population relative to that earned by the poorest 20%. Thus, it can be expressed using quantile shares, where a quantile share is the share of total income earned by all of the units up to a given quantile. The aim of this paper is to propose an improved methodology for the estimation and variance estimation of the QSR in a complex sampling design framework. Because the QSR is a non-linear function of interest, the estimation of its sampling variance requires advanced methodology. Moreover, a non-trivial obstacle in the estimation of quantile shares in finite populations is the non-unique definition of a quantile. Thus, two different conceptions of the quantile share are presented in the paper, leading us to two different estimators of the QSR. Regarding variance estimation, Osier (2006, 2009) proposed a variance estimator based on linearization techniques. However, his method involves Gaussian kernel smoothing of cumulative distribution functions. Our approach, also based on linearization, shows that no smoothing is needed. The construction of confidence intervals is discussed and a proposition is made to account for the skewness of the sampling distribution of the QSR. Finally, simulation studies are run to assess the relevance of our theoretical results.

Keywords:
Inequality measure
Sampling
Variance
Estimation
Quantile
Confidence intervals

1. Introduction

Nowadays, the Quintile Share Ratio (QSR) is a widely used measure of inequality. Together with the Gini index, it is one of the two Laken indicators of inequality selected at the European Council in Laken, Belgium in 2001. Laken indicators are used in the European Statistics on Income and Living Conditions (EU-SILC) program run by Eurostat (Eurostat, 2005; Traat, 2006)). The QSR is a function of quantile shares, where a quantile share is the share of total income earned by all of the units up to a given quantile.

This paper focuses on conducting statistical inference for the QSR in a complex random sampling framework. However, the proposed method can be applied to other quantile share-based measures. Because the QSR is a non-linear function of the incomes, variance estimation is not straightforward and requires specific techniques. The variance estimators proposed here are based on the linearization approach by Deville (1999). Inference for the QSR using this approach has already been conducted by Osier (2006, 2009) and similar work has been done for the Gini index (Deville, 1996, 1999; Berger, 2008;

* Corresponding author. Tel.: +41 32 718 13 54.
E-mail address: matti.langel@unine.ch (M. Langel).

Barrett and Donald, 2009). However, Osier's approach is intricate because it requires kernel smoothing of cumulative distribution functions. With the improvement proposed in this paper, smoothing is no longer required. Moreover, an alternative estimator and variance estimator are presented, and a set of simulations advocate in favor of the latter.

The paper is organized as follows. Section 2 starts with a presentation of key concepts, namely quantiles, quantile shares and partial sums. The continuous case is discussed to begin with, but emphasis is placed on finite population expressions as well as on estimators under complex sampling designs. In this section, it is also stressed that the partial sum centralizes the main issues in conducting valid inference for the QSR. Thus, two distinct finite population expressions and estimators of the partial sum are presented. The first one is based on quantiles and leads to a natural expression of the QSR, while the second is an alternative expression that gets around the finite population quantile issue and leads to another finite population expression of the inequality measure of interest. Both are described in Section 3 and their respective estimators are given.

Section 4 is a succinct description of the linearization technique using influence functions for variance estimation, as initially proposed by Deville (1999). The approach is then applied in parallel to both estimators of the QSR, providing us with two distinct variance estimators. Firstly, the influence functions of both expressions of the partial sum are derived in Sections 5 and 6. In these sections, we point out that, unlike in the approach by Osier (2006, 2009), no smoothing is needed. The two resulting variance estimators are then derived in Sections 7 and 8. A discussion on confidence intervals and skewness issues is proposed in Section 9. Finally, two sets of simulations on real data are presented in Section 10, preceding some concluding remarks.

2. Estimation of quantile shares

Quantile share-based measures constitute a very interesting class of inequality indices in their capacity to detect perturbations at different levels of an income distribution (Langel and Tillé, 2009). However, inference on these measures is not straightforward, especially when dealing with complex sampling designs. Consider a continuous strictly increasing cumulative distribution function $F(y)$ and $F'(y)$, its derivative and probability density function. Also, let us denote Q_α, the quantile of order α, such that $F(Q_\alpha)=\alpha$. The quantile function can be written as the inverse of the cumulative distribution function $Q_\alpha = F^{-1}(\alpha)$. A quantile share is the share of total income earned by all the income earners up to quantile of order α. The definition for the continuous case is

$$L(\alpha)=\frac{\int_0^{Q_\alpha} u\, dF(u)}{\int_0^{\infty} u\, dF(u)}.$$

This expression is also frequently referred to as the Lorenz function or Lorenz curve (Lorenz, 1905; Gastwirth, 1972; Cowell, 1977; Kovacevic and Binder, 1997), which is a central tool of inequality theory.

Let U denote a finite population of N identifiable units $u_1,\ldots,u_k,\ldots,u_N$. For the sake of simplicity, we will hereafter denote unit u_k by its identifier k. Associated with each unit k is the value y_k of some characteristic of interest, for example income. To lighten the notation, we will assume with no loss of generality that all y_k's are distinct and sorted. The finite population quantile share is $L(\alpha)=Y_\alpha/Y$, where $Y=\sum_{k\in U} y_k$ and

$$Y_\alpha=\sum_{k\in U} y_k \mathbb{1}[\mathrm{y_k}\le \mathrm{Q}_\alpha], \tag{2.1}$$

with Q_α, the quantile of order α and with $\mathbb{1}(\mathrm{A})=1$ if A is true and 0 otherwise. Expression (2.1) is thereafter denoted as the *partial sum* of income y. The finite population quantile share $L(\alpha)$ is the cumulative sum of income up to a given quantile Q_α over the total income. Or, in other words, the share of total income earned by the αN poorer units. In the following, we will mainly focus on the partial sum Y_α, because it embodies the complex part of the quantile share.

A classical notation from survey sampling theory is used hereafter. Thus, let us denote S, a random sample of size n, and the function $p(s)=\Pr(S=s)$, which gives the probability of selecting the particular sample $s\subset U$. The inclusion probability of unit k is denoted π_k and defined such that $\pi_k=\Pr(k\in S)$. Also, w_k stands for the weight of unit k. Weights can simply be the inverse of the inclusion probability $w_k=1/\pi_k$, but can also result from a calibration procedure (Deville and Särndal, 1992) or non-response adjustments (Särndal and Lundström, 2005). In the following, it is assumed that the sampling design used to draw sample S is associated with a known expression of the variance of the estimated total

$$\widehat{Y}=\sum_{k\in S} w_k y_k. \tag{2.2}$$

From a sample, Y_α can be estimated in a similar fashion using the plug-in estimator

$$\widehat{Y}_\alpha=\sum_{k\in S} w_k y_k \mathbb{1}[\mathrm{y_k}\le \widehat{\mathrm{Q}}_\alpha], \tag{2.3}$$

where $\widehat{Q}_\alpha$ is an estimator of quantile Q_α. Both the quantile and its estimator have to be precisely defined. While the definition of a quantile in a continuous distribution is clear and unique, it is not so in the finite population context, where F, the cumulative distribution of income y, is a step function. Accordingly, obtaining a univocal definition of quantile Q_α is not possible. In the paper by Hyndman and Fan (1996), nine different definitions of sample quantiles are described, all of

them existing in the literature and in statistical packages. In order to compute $\widehat{Y}_\alpha$ in the simulation study (Section 10), we will be using the fourth definition of the quantile of Hyndman and Fan (1996), which is based on a simple linear interpolation of the cumulative distribution function:

$$Q_\alpha = y_{k-1} + (y_k - y_{k-1})[\alpha N - (k-1)], \tag{2.4}$$

where $\alpha N < k \leq \alpha N + 1$. The quantile can be estimated from a sample by

$$\widehat{Q}_\alpha = y_{k-1} + (y_k - y_{k-1}) \left(\frac{\alpha \widehat{N} - W_{k-1}}{w_k} \right),$$

where $W_k = \sum_{\ell \in S} w_\ell \mathbb{1}[y_\ell \leq y_k]$, $\widehat{N} = W_n$ and the value of k is such that $W_{k-1} < \alpha \widehat{N} \leq W_k$.

As emphasized above, the value of Q_α, and consequently of Y_α, is dependent on how the discontinuities of the cumulative distribution function are dealt with. This issue fosters the use of another definition of the partial sum that is not directly dependent on the definition of the quantile:

$$\widetilde{Y}_\alpha = \sum_{k \in U} y_k H[\alpha N - (k-1)], \tag{2.5}$$

where

$$H(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases} \tag{2.6}$$

Here H is the cumulative distribution function of a uniform random variable. Under this definition, $\widetilde{Y}_\alpha$ is a strictly increasing function of α. This is a desirable property for the estimation of the quantile share and its variance in sampling from a finite population. Partial sum $\widetilde{Y}_\alpha$ can be estimated from a sample S using the plug-in estimator

$$\widehat{\widetilde{Y}}_\alpha = \sum_{k \in S} w_k y_k H \left(\frac{\alpha \widehat{N} - W_{k-1}}{w_k} \right). \tag{2.7}$$

In the following, both Y_α (2.1) and $\widetilde{Y}_\alpha$ (2.5) will be considered in order to define the QSR and provide variance estimators using influence functions.

3. The quintile share ratio

The QSR is defined as the ratio of the total income earned by the richest 20% of the population relative to that earned by the poorest 20%. For the continuous case, we thus have

$$\mathrm{QSR} = \frac{1 - L(0.8)}{L(0.2)}.$$

In finite populations, the QSR can be viewed as a function of quantile shares or as a function of partial sums. Because we have proposed two different finite population definitions of the partial sum, the QSR can be defined by

$$\mathrm{QSR} = \frac{Y - Y_{0.8}}{Y_{0.2}}, \tag{3.1}$$

or by

$$\widetilde{\mathrm{QSR}} = \frac{Y - \widetilde{Y}_{0.8}}{\widetilde{Y}_{0.2}}, \tag{3.2}$$

and can be, respectively, estimated from a sample by

$$\widehat{\mathrm{QSR}} = \frac{\widehat{Y} - \widehat{Y}_{0.8}}{\widehat{Y}_{0.2}}, \tag{3.3}$$

or by

$$\widehat{\widetilde{\mathrm{QSR}}} = \frac{\widehat{Y} - \widehat{\widetilde{Y}}_{0.8}}{\widehat{\widetilde{Y}}_{0.2}}. \tag{3.4}$$

4. Approximation of the variance by linearization

The estimators of the Quintile Share Ratio as defined in (3.3) and (3.4) are non-linear statistics, and, therefore, a general expression for their sampling variance is not known. In the literature, a variety of methods such as resampling techniques

or linearization allow for variance estimation of complex statistics (for a survey of most existing methods, see Wolter, 2007). Linearization methods have given rise to a lot of research using different approaches (Woodruff, 1971; Binder and Patak, 1994; Kovacevic and Binder, 1997; Deville, 1999; Demnati and Rao, 2004). This paper focuses on the linearization method developed by Deville (1999). His approach is based on the influence function, a predominant notion in the field of robust statistics (Hampel et al., 1985). The idea behind this method is to study the influence of unit k on the population parameter of interest by adding an infinitesimal variation of the weight to this unit. A population parameter θ can be written as a functional $T(M)$, where measure M allocates a unit mass to all $k \in U$. The influence function of T is then defined as

$$z_k = I[T(M)]_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t},$$

where δ_k denotes the Dirac measure for unit k. The term *linearized variable* is used hereafter to denote z_k. This terminology is used by Deville (1999) and advocated by Skinner (2004). Under asymptotic conditions described in Deville (1999), it is shown that the variance of the estimated total of the linearized variable z_k is an approximation to the variance of statistic $\widehat{\theta}$ (an estimator of θ):

$$\text{var}\left(\sum_{k \in S} z_k w_k\right) \approx \text{var}(\widehat{\theta}). \tag{4.1}$$

In practice, values z_k at the left-hand side of Expression (4.1) are not known because they rely on unavailable information at the population level. Thus, the z_k's are estimated from the sample by using the plug-in estimator $\widehat{z}_k = I[T(\widehat{M})]_k$, where $\widehat{M}$ is the measure allocating a mass w_k to all $k \in S$. With the proposed method, the variance of a complex statistic $\widehat{\theta}$ can be estimated under any sampling design for which the expression of the variance of the estimator of a total is available.

5. Linearization of Y_α

The method described in the above section can be applied to obtain an expression for the influence function of both definitions of the partial sum. Derivation for the influence function of Y_α is shown in the present section, whereas Section 6 presents derivation for $\widetilde{Y}_\alpha$.

For Y_α, a solution is proposed in Osier (2006, 2009):

$$I(Y_\alpha)_k = y_k \mathbb{1}[y_k \le Q_\alpha] + \widetilde{S}'(Q_\alpha) I(Q_\alpha)_k, \tag{5.1}$$

where $\widetilde{S}'$ is the derivative of $\widetilde{S}$, a smoothed function of

$$S(y) = \sum_{k \in U} y_k \mathbb{1}[y_k \le y].$$

Two issues arise from this expression: the smoothing of S and the computation of $I(Q_\alpha)_k$, the influence function of quantile Q_α. A solution to the latter issue is proposed by Deville (1999):

$$I(Q_\alpha)_k = -\frac{\mathbb{1}[y_k \le Q_\alpha] - \alpha}{\widetilde{F}'(Q_\alpha) N}, \tag{5.2}$$

with $\widetilde{F}'(y)$, the derivative of $\widetilde{F}(y)$, a smoothed function of

$$F(y) = \frac{1}{N} \sum_{k \in U} \mathbb{1}[y_k \le y].$$

At this point, computing $I(Y_\alpha)_k$ thus implies the smoothing of two discontinuous step functions, S and F. Deville (1999) suggests kernel smoothing for F. In order to estimate the variance of the QSR estimated from a sample for various European Union member states, Osier (2006, 2009) has applied (5.1) using Gaussian kernel smoothing of S and F. We propose hereafter a simpler solution that does not require smoothing of the latter functions.

Let us momentarily consider $\widetilde{S}(y)$ and $\widetilde{F}(y)$, the smoothed functions of S and F, respectively. Let also $G(y) = \widetilde{S}(y)/Y$. Since $\widetilde{S}(y)$ is differentiable, $G(y)$ is also differentiable. If $G'(y)$ denotes the derivative of $G(y)$, then

$$\widetilde{S}'(y) = G'(y) Y. \tag{5.3}$$

Functions $G(y)$ and $F(y)$ are both cumulative distribution functions. $G(y)$ can also be defined by

$$G(y) = \frac{\int_0^y u \, d\widetilde{F}(u)}{\int_0^\infty u \, d\widetilde{F}(u)} = \frac{N \int_0^y u \, d\widetilde{F}(u)}{Y}.$$

Thus, $G'(y)$, can be written

$$G'(y) = \frac{Ny\widetilde{F}'(y)}{Y}. \tag{5.4}$$

Let us now substitute (5.2) and (5.3) in Expression (5.1):

$$I(Y_\alpha)_k = y_k \mathbb{1}[y_k \le Q_\alpha] - \frac{Y}{N}\frac{G'(Q_\alpha)}{\widetilde{F}'(Q_\alpha)}(\mathbb{1}[y_k \le Q_\alpha] - \alpha). \tag{5.5}$$

Moreover, from (5.4) we have

$$\frac{G'(Q_\alpha)}{\widetilde{F}'(Q_\alpha)} = \frac{Q_\alpha N}{Y}, \tag{5.6}$$

and thus, replacing (5.6) into (5.5), we finally obtain

$$I(Y_\alpha)_k = \alpha Q_\alpha - (Q_\alpha - y_k)\mathbb{1}[y_k \le Q_\alpha], \tag{5.7}$$

where no smoothing is needed. Indeed, densities $\widetilde{F}'$ and G' do not appear in Result (5.7), making the computation of the influence function of Y_α markedly more straightforward than the method initially proposed by Osier (2006, 2009).

6. Linearization of $\widetilde{Y}_\alpha$

The influence function of $\widetilde{Y}_\alpha$ can also be derived. First, the rule for the linearization of a product (Deville, 1999; Dell et al., 2002) is applied to Eq. (2.5)

$$I(\widetilde{Y}_\alpha)_k = y_k H(\alpha N - k + 1) + \sum_{j \in U} y_j I[H(\alpha N - j + 1)]_k. \tag{6.1}$$

The influence function of H is

$$I[H(\alpha N - j + 1)]_k = \mathbb{1}(0 < \alpha N - j + 1 \le 1)[\alpha - \mathbb{1}(k < j)],$$

and because $\mathbb{1}[k < j] = \mathbb{1}[y_k < y_j]$, we obtain

$$\sum_{j \in U} y_j I[H(\alpha N - j + 1)]_k = [\alpha - \mathbb{1}(y_k < \widetilde{Q}_\alpha)]\widetilde{Q}_\alpha, \tag{6.2}$$

where $\widetilde{Q}_\alpha$ denotes the first definition of the finite population quantile in the paper by Hyndman and Fan (1996):

$$\widetilde{Q}_\alpha = y_i \quad \text{where } i - 1 < \alpha N \le i.$$

Finally, the influence function of $\widetilde{Y}_\alpha$ is obtained by substituting (6.2) in (6.1):

$$I(\widetilde{Y}_\alpha)_k = y_k H(\alpha N - k + 1) + [\alpha - \mathbb{1}(y_k < \widetilde{Q}_\alpha)]\widetilde{Q}_\alpha. \tag{6.3}$$

No prior definition of the quantile in finite populations was required in order to derive $I(\widetilde{Y}_\alpha)_k$. However, quantile $\widetilde{Q}_\alpha$ appears in the final expression.

7. Linearization of QSR

The influence function of QSR is computed by applying the derivation rule for the linearization of a ratio (Deville, 1999; Dell et al., 2002) on Expression (3.1),

$$I(\mathrm{QSR})_k = \frac{y_k - I(Y_{0.8})_k}{Y_{0.2}} - \frac{(Y - Y_{0.8})I(Y_{0.2})_k}{Y_{0.2}^2},$$

and by replacing $I(Y_{0.2})_k$ and $I(Y_{0.8})_k$ with the result obtained in (5.7). We, therefore, have

$$I(\mathrm{QSR})_k = \frac{y_k - \{0.8Q_{0.8} - (Q_{0.8} - y_k)\mathbb{1}[y_k \le Q_{0.8}]\}}{Y_{0.2}} - \frac{(Y - Y_{0.8})\{0.2Q_{0.2} - (Q_{0.2} - y_k)\mathbb{1}[y_k \le Q_{0.2}]\}}{Y_{0.2}^2}.$$

Our final aim here is to derive a sampling variance estimator for $\widehat{\mathrm{QSR}}$. Linearization theory shows us that, with $z_k = I(\mathrm{QSR})_k$,

$$\mathrm{var}(\widehat{\mathrm{QSR}}) \approx \mathrm{var}\left(\sum_{k \in S} z_k w_k\right).$$

In practice, however, the linearized variable z_k involves unavailable information at the population level and has to be estimated from the sample by its plug-in estimator

$$\widehat{z}_k = \frac{y_k - \{0.8\widehat{Q}_{0.8} - (\widehat{Q}_{0.8} - y_k)\mathbb{1}[y_k \le \widehat{Q}_{0.8}]\}}{\widehat{Y}_{0.2}} - \frac{(\widehat{Y} - \widehat{Y}_{0.8})\{0.2\widehat{Q}_{0.2} - (\widehat{Q}_{0.2} - y_k)\mathbb{1}[y_k \le \widehat{Q}_{0.2}]\}}{\widehat{Y}_{0.2}^2}.$$

The estimated variance of $\widehat{\mathrm{QSR}}$ is obtained by estimating the variance of the weighted sum $\widehat{Z} = \sum_{k \in S} \widehat{z}_k w_k$. The method is thus easily applicable to a whole variety of complex sampling designs. Under a simple random sampling design without

replacement, the variance estimator for $\widehat{\mathrm{QSR}}$ is

$$\widehat{\mathrm{var}}_{lin}(\widehat{\mathrm{QSR}}) = \frac{N(N-n)}{n(n-1)} \sum_{k \in S} (\widehat{z}_k - \overline{z})^2, \tag{7.1}$$

with $\overline{z} = n^{-1} \sum_{k \in S} \widehat{z}_k$. In the following, the latter estimator is always referred to as $\widehat{\mathrm{var}}_{lin}(\widehat{\mathrm{QSR}})$ to emphasize the fact that it is obtained through a linearization technique and to clearly distinguish it from the Monte Carlo estimator used in the simulation studies.

8. Linearization of $\widetilde{\mathrm{QSR}}$

A very similar derivation can be done for the influence function of $\widetilde{\mathrm{QSR}}$. Indeed, we have

$$I(\widetilde{\mathrm{QSR}})_k = \frac{y_k - I(\widetilde{Y}_{0.8})_k}{\widetilde{Y}_{0.2}} - \frac{(Y - \widetilde{Y}_{0.8}) I(\widetilde{Y}_{0.2})_k}{\widetilde{Y}_{0.2}^2},$$

and applying (6.3), $I(\widetilde{\mathrm{QSR}})_k$ can be rewritten

$$I(\widetilde{\mathrm{QSR}})_k = \frac{y_k - \{y_k M(0.8N - k + 1) + \widetilde{Q}_{0.8}[0.8 - \mathbb{1}(\mathrm{y_k} < \widetilde{Q}_{0.8})]\}}{\widetilde{Y}_{0.2}} - \frac{(Y - \widetilde{Y}_{0.8})\{y_k M(0.2N - k + 1) + \widetilde{Q}_{0.2}[0.2 - \mathbb{1}(\mathrm{y_k} < \widetilde{Q}_{0.2})]\}}{\widetilde{Y}_{0.2}^2}.$$

With $\widetilde{z}_k = I(\widetilde{\mathrm{QSR}})_k$, we have

$$\mathrm{var}(\widehat{\widetilde{\mathrm{QSR}}}) \approx \mathrm{var}\left(\sum_{k \in S} \widetilde{z}_k w_k \right),$$

and in practice, the (unknown) $\widetilde{z}_k$'s are replaced by the plug-in estimator

$$\widehat{\widetilde{z}}_k = \frac{y_k - \left\{ y_k M \left(\frac{0.8\widehat{N} - W_{k-1}}{w_k} \right) + \widehat{\widetilde{Q}}_{0.8}[0.8 - \mathbb{1}(\mathrm{y_k} < \widehat{\widetilde{Q}}_{0.8})] \right\}}{\widehat{\widetilde{Y}}_{0.2}} - \frac{(\widehat{Y} - \widehat{\widetilde{Y}}_{0.8}) \left\{ y_k M \left(\frac{0.2\widehat{N} - W_{k-1}}{w_k} \right) + \widehat{\widetilde{Q}}_{0.2}[0.2 - \mathbb{1}(\mathrm{y_k} < \widehat{\widetilde{Q}}_{0.2})] \right\}}{\widehat{\widetilde{Y}}_{0.2}^2},$$

where $\widehat{\widetilde{Q}}_{\alpha} = y_i$, with $W_{i-1} < \alpha N \leq W_i$. Finally, the variance estimator for a simple random sampling design without replacement is constructed similarly as in (7.1)

$$\widehat{\mathrm{var}}_{lin}(\widehat{\widetilde{\mathrm{QSR}}}) = \frac{N(N-n)}{n(n-1)} \sum_{k \in S} (\widehat{\widetilde{z}}_k - \overline{\widetilde{z}})^2, \tag{8.1}$$

where $\overline{\widetilde{z}} = n^{-1} \sum_{k \in S} \widehat{\widetilde{z}}_k$.

9. Construction of a confidence interval

As shown by the simulation results in Section 10 below, the variances are successfully and accurately estimated by using the linearization method. However, because of the skewness of the sampling distributions of the statistics $\widehat{\mathrm{QSR}}$ and $\widehat{\widetilde{\mathrm{QSR}}}$, the normality-based confidence intervals built around these estimators are somewhat less convincing.

To account for the skewness issues in the interval estimation of $\widehat{\mathrm{QSR}}$ and $\widehat{\widetilde{\mathrm{QSR}}}$, we propose an alternative method for the construction of a more reliable confidence interval. The method, based on Box–Cox transformations (Box and Cox, 1964), aims to obtain a less skewed distribution after transformation, and consequently to build confidence intervals on the latter distribution. The Box–Cox transformations for a parameter θ are given by

$$\theta^{(\lambda)} = \begin{cases} \dfrac{\theta^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log\theta & \text{if } \lambda = 0. \end{cases} \tag{9.1}$$

The method consists of constructing confidence intervals for $\mathrm{QSR}^{(\lambda)}$ and $\widehat{\widetilde{\mathrm{QSR}}}^{(\lambda)}$. For that purpose, variance estimators of the latter statistics also need to be derived. With the linearized variable $z_k = I(\mathrm{QSR})_k$, the influence function of any Box–Cox transformation $\mathrm{QSR}^{(\lambda)}$ is

$$I[\mathrm{QSR}^{(\lambda)}]_k = z_k \mathrm{QSR}^{\lambda - 1}. \tag{9.2}$$

Thus, we also have

$$I[\widehat{\mathrm{QSR}}^{(\lambda)}]_k = \widehat{z}_k \widehat{\mathrm{QSR}}^{\lambda - 1}, \tag{9.3}$$

and the variance estimator

$$\widehat{\mathrm{var}}_{lin}[\widehat{\mathrm{QSR}}^{(\lambda)}] = \widehat{\mathrm{QSR}}^{2(\lambda - 1)} \widehat{\mathrm{var}}_{lin}(\widehat{\mathrm{QSR}}). \tag{9.4}$$

Expression (9.4) shows that a variance estimator and confidence interval can be directly derived for any value of parameter λ. An alternative confidence interval for the untransformed QSR can thus be obtained by computing the inverse transformation of the lower and upper bounds of the confidence interval for $QSR^{(\lambda)}$. The procedure can be applied equivalently to $\widehat{QSR}^{(\lambda)}$. The aim is to choose the value of λ which yields the most symmetric distribution. The sampling distribution of the statistic is unknown. However, our simulation studies below (Section 10) show that $\lambda = -1$ seems to be an appropriate solution.

10. Simulation studies

Two simulation studies have been carried out on real data to evaluate the quality of the linearization variance estimators of the QSR proposed in this paper. The data used in the simulations is the household taxable income of the Canton of Neuchâtel, Switzerland for year 2007. It regroups a population of $N=88{,}106$ non-null income earners. The data is highly positively skewed, which is not surprising for income data.

The first simulation is dedicated to $\widehat{QSR}$, while the second focuses on $\widehat{\widehat{QSR}}$. For the first simulation study, 100,000 samples of size $n=1000$ are drawn using a simple random sampling design without replacement. Firstly, the value of $\widehat{QSR}$ is computed on each sample, which provides us with $var_{sim}(\widehat{QSR})$, a Monte Carlo estimator of the variance of $\widehat{QSR}$ under the simulations. The linearization variance estimator $\widehat{var}_{lin}(\widehat{QSR})$ (Eq. (7.1)) is then computed on each sample using the linearized variable $\hat{z}$. The main goal of this study is to analyze the quality of the latter estimator in terms of bias and variance with respect to $var_{sim}(\widehat{QSR})$. For this purpose, the Monte Carlo expected value and variance of the linearization variance estimators, respectively, denoted $E_{sim}[\widehat{var}_{lin}(\widehat{QSR})]$ and $var_{sim}[\widehat{var}_{lin}(\widehat{QSR})]$, are computed.

Likewise, the second simulation study aims to compare $var_{sim}(\widehat{\widehat{QSR}})$ and $\widehat{var}_{lin}(\widehat{\widehat{QSR}})$. The method and sampling designs are exactly identical in both studies. Eventually, the joint analysis of the results from the two simulations allows for a brief comparison of both definitions of the Quintile Share Ratio, and of their capacity to provide a reliable estimation. The results can be viewed in Table 1.

The relative bias for $\widehat{\theta}$ and $\widehat{var}_{lin}(\widehat{\theta})$ are, respectively, defined by

$$RB(\widehat{\theta}) = [E_{sim}(\widehat{\theta}) - \theta]/\theta,$$

and

$$RB[\widehat{var}_{lin}(\widehat{\theta})] = \{E_{sim}[\widehat{var}_{lin}(\widehat{\theta})] - var_{sim}(\widehat{\theta})\}/var_{sim}(\widehat{\theta}).$$

The two definitions above are slightly different because while $\widehat{\theta}$ is compared to the true finite population value θ, the variance estimator is compared to the Monte Carlo estimator $var_{sim}(\widehat{\theta})$ which approximates the true value.

The results show that the variance estimators obtained via the linearization method are very close to the Monte Carlo variance estimator. This is emphasized by the relative bias of the linearization variance estimators in both simulations (-0.68% and -0.07%, respectively). Point estimation is also accurate with a relative bias of -0.60% for the estimator $\widehat{QSR}$ and of 0.12% for $\widehat{\widehat{QSR}}$. Although these results show the validity and accuracy of both categories of estimators, they also emphasize that $\widehat{\widehat{QSR}}$ leads to a slightly better inference for point and variance estimation than $\widehat{QSR}$.

It is also to be noticed that 95% confidence intervals in both simulations are only partially satisfactory in terms of coverage rate (CR). The coverage rate of the confidence intervals constructed around $\widehat{\widehat{QSR}}$ is, however, distinctly better than for $\widehat{QSR}$ with 93.1% and 92.0%, respectively. This is one of the major reasons why inference on $\widehat{\widehat{QSR}}$ should be preferred. Also, a possible improvement of the coverage rate is discussed below.

Table 1
Results for both simulation studies (100,000 replications each). The middle column summarizes results from the simulation study for which estimator $\widehat{QSR}$ and linearized variable $\hat{z}$ are used. Respectively, results displayed on the right-hand side for the second simulation study were computed using $\widehat{\widehat{QSR}}$ and $\hat{\hat{z}}$. The coverage rate for a 95% normality based confidence interval for QSR is denoted CR.

	Simulations on $\widehat{\theta} = \widehat{QSR}$	**Simulations on $\widehat{\theta} = \widehat{\widehat{QSR}}$**
$var_{sim}(\widehat{\theta})$	0.9216	0.9280
$E_{sim}[\widehat{var}_{lin}(\widehat{\theta})]$	0.9153	0.9273
$var_{sim}[\widehat{var}_{lin}(\widehat{\theta})]$	1.7731	1.7793
$RB[\widehat{var}_{lin}(\widehat{\theta})]$	-0.68%	-0.07%
$RB(\widehat{\theta})$	-0.60%	0.12%
CR	92.0%	93.1%

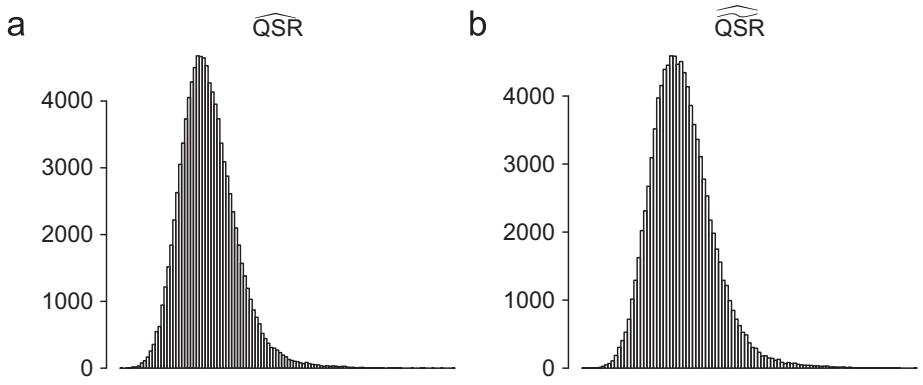

Fig. 1. Histograms of the distributions of $\widehat{QSR}$ and $\widehat{\widehat{QSR}}$ computed on 100,000 simple random samples without replacement of size $n=1000$.

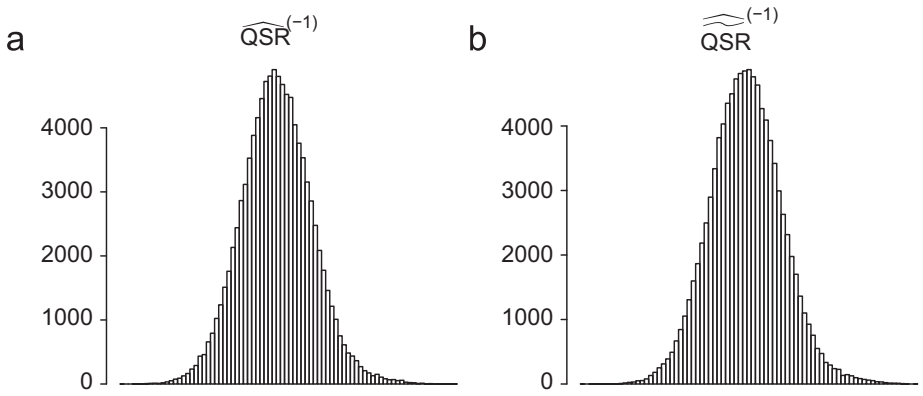

Fig. 2. Histograms of the distributions of $\widehat{QSR}^{(-1)}$ and $\widehat{\widehat{QSR}}^{(-1)}$ computed on 100,000 simple random samples without replacement of size $n=1000$.

As shown in Fig. 1, a substantial level of skewness has been observed in the sampling distributions of $\widehat{QSR}$ and $\widehat{\widehat{QSR}}$. This seems to result from the skewness of the incomes and from the sensitivity of these statistics to extreme values. The skewness in the distributions of $\widehat{QSR}$ and $\widehat{\widehat{QSR}}$ makes it difficult for us to achieve reliable confidence intervals. To account for this question, we have applied the procedure proposed in Section 9, and have constructed confidence intervals for two different Box–Cox transformations of $\widehat{QSR}$ and $\widehat{\widehat{QSR}}$: $\lambda=0$ and -1. Fig. 2 displays the improvement in terms of skewness of the distribution provided by the Box–Cox $\lambda=-1$ transformation.

In order to produce confidence intervals for these transformations, linearization variance estimators for $\lambda=0$ and -1 are obtained from Expression (9.4):

$$\widehat{var}_{lin}[\widehat{QSR}^{(0)}]=\frac{1}{\widehat{QSR}^2}\widehat{var}_{lin}(\widehat{QSR}),$$

$$\widehat{var}_{lin}[\widehat{QSR}^{(-1)}]=\frac{1}{\widehat{QSR}^4}\widehat{var}_{lin}(\widehat{QSR}),$$

and similarly with $\widehat{\widehat{QSR}}$ for the second set of simulations. A confidence interval for QSR is then simply obtained by applying a back transformation on the confidence bounds. As shown in Table 2, the transformations result in a substantial improvement of the coverage rate for both simulation studies. Because they performed a more severe correction of the asymmetry, the $\lambda=-1$ transformations yield better results than the log-transformation ($\lambda=0$), with coverage rates of 93.6% in the first study, and 94.0% in the second.

11. Conclusion

The goal of this paper was to provide reliable tools for finite population inference for the Quintile Share Ratio. We have proposed two distinct estimators for the latter inequality measure, as well as two corresponding variance estimators. Although presented in the simulation studies for simple random sampling, all proposed estimators can be used with

Table 2
Coverage rates (CR) for 95% confidence intervals for QSR using Box–Cox transformations. The initial coverage rate (untransformed) from Table 1 is reprinted for comparison purposes.

	Simulations on $\widehat{\mathrm{QSR}}$ (%)	**Simulations on** $\widehat{\widehat{\mathrm{QSR}}}$ (%)
CR (untransformed)	92.0	93.1
CR ($\lambda = 0$)	92.9	93.8
CR ($\lambda = -1$)	93.6	94.0

complex sampling designs. Indeed, because it focuses on linearization techniques, the method holds for any sampling design as long as the expression for the variance of an estimated total is known.

Next, the two variance estimators proposed in this paper do not require the smoothing of any cumulative distribution function or density. Accordingly, variance estimation for the Quintile Share Ratio is not only faster and simpler with our technique than with past methodology, but it also avoids issues that are inherent to non-parametric smoothing, such as the choice of the type of kernel or the size of the bandwidth. This is thus a sensible improvement to the method proposed by Osier (2006, 2009).

Also, while the paper focuses on the Quintile Share Ratio, the main contribution involves more specifically the partial sum and the quantile share. Thus, the method is not restricted to the Quintile Share Ratio and can be applied to other quantile share-based functions of interest. Our simulation studies on real income data confirm the theoretical findings and show that the method is accurate and straightforward to apply. As in depth analysis of the simulations shows that estimating the quintile share ratio with $\widehat{\widehat{\mathrm{QSR}}}$ seems to be slightly more favorable in terms of bias and coverage rate than with $\widehat{\mathrm{QSR}}$. Using real data also reminds us that skewness issues as well as the sensitivity of the statistic to extreme values are obstacles to reliable inference. We have shown that Box–Cox transformations can help address the problem of skewness. In particular, studying the $\lambda = -1$ transformation instead of the Quintile Share Ratio itself can be a valuable alternative. Finally, studies on the robustness of inequality measures (Hulliger and Munnich, 2006) can provide insights on the issue of sensitivity to outliers.

Acknowledgments

The authors are grateful to the Office Cantonal de la Statistique (Canton de Neuchâtel, Switzerland) and especially to Gérard Geiser for the dataset. The authors would also like to thank an anonymous referee for useful comments and suggestions. This research is supported by Grant no. 200021-121604 of the Swiss National Science Foundation.

References

Barrett, G.F., Donald, S.G., 2009. Statistical inference with generalized Gini indices of inequality, poverty, and welfare. Journal of Business and Economic Statistics 27 (1), 1–17.
Berger, Y.G., 2008. A note on asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. Journal of Official Statistics 24, 541–555.
Binder, D.A., Patak, Z., 1994. Use of estimating functions for estimation from complex surveys. Journal of the American Statistical Association 89, 1035–1043.
Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. (With discussion). Journal of the Royal Statistical Society, Series B. Methodological 26 (2), 211–252.
Cowell, F.A., 1977. Measuring Inequality. Philip Allan, Oxford.
Dell, F., d'Haultfoeuille, X., Fevrier, P., Masse, E., 2002. Mise en oeuvre du calcul de variance par linéarisation. INSEE-Méthodes: Actes des Journées de Méthodologie Statistique, pp. 73–104.
Demnati, A., Rao, J.N.K., 2004. Linearization variance estimators for survey data (with discussion). Survey Methodology 30, 17–34.
Deville, J.-C., 1996. Estimation de la variance du coefficient de Gini estimé par sondage. Actes des journées de Méthodologie Statistique, INSEE 69-70-71, 269–288.
Deville, J.-C., 1999. Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology 25, 193–204.
Deville, J.-C., Särndal, C.-E., 1992. Calibration estimators in survey sampling. Journal of the American Statistical Association 87, 376–382.
Eurostat, 2005. The continuity of indicators during the transition between ECHP and EU-SILC. Technical Report, Working Papers and Studies, Office for Official Publications of the European Communities, Luxembourg.
Gastwirth, J.L., 1972. The estimation of the Lorenz curve and Gini index. Review of Economics and Statistics 54, 306–316.
Hampel, F.R., Ronchetti, E., Rousseuw, P.J., Stahel, W., 1985. Robust Statistics: The Approach Based on the Influence Function. Wiley, New-York.
Hulliger, B., Munnich, R., 2006. Variance estimation for complex surveys in the presence of outliers. In: Proceedings of the Section on Survey Research Methods. American Statistical Association, pp. 3153–3161.
Hyndman, R.J., Fan, Y., 1996. Sample quantiles in statistical packages. American Statistician 50, 361–365.
Kovacevic, M.S., Binder, D.A., 1997. Variance estimation for measures of income inequality and polarization—the estimating equations approach. Journal of Official Statistics 13, 41–58.
Langel, M., Tillé, Y., 2009. An evaluation of the performance of inequality measures for the detection of changes in an income distribution. Technical Report, University of Neuchatel.
Lorenz, M., 1905. Methods of measuring the concentration of wealth. Publications of the American Statistical Association 9, 209–219.
Osier, G., 2006. Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation. Technical Report, Eurostat and Statistics Finland Methodological Workshop on EU-SILC, Helsinki, 7–8 November 2006.
Osier, G., 2009. Variance estimation for complex indicators of poverty and inequality using linearization techniques. Survey Research Methods 3, 167–195.

Särndal, C.-E., Lundström, S., 2005. Estimation in Surveys with Nonresponse. Wiley, New York.
Skinner, C.J., 2004. Comment on Demnati and Rao: linearization variance estimators for survey data. Survey Methodology 30, 17–18.
Traat, I., 2006. Variance of quantile estimators in household surveys. In: Workshop on Survey Sampling Theory and Methodology. Central Statistical Bureau of Latvia, Ventspils, pp. 62–65.
Wolter, K.M., 2007. Introduction to Variance Estimation, second ed. Springer, New York.
Woodruff, R.S., 1971. A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association 66, 411–414.