LIS
Working Paper Series

No. 601

Fast Methods for Jackknifing Inequality Indices

Lynn A. Karoly and Carsten Schröder

December 2013

Luxembourg Income Study (LIS), asbl

Fast Methods for Jackknifing Inequality Indices

Lynn A. Karoly, RAND

Carsten Schröder[♠], University of Kiel, Department of Economics

Abstract. The jackknife is a resampling method that uses subsets of the original database by leaving out one observation at a time from the sample. The paper outlines a procedure to obtain jackknife estimates for several inequality indices with only a few passes through the data. The number of passes is independent of the number of observations. Hence, the method provides an efficient way to obtain standard errors of the estimators even if sample size is large. We apply our method using micro data on individual incomes for Germany and the US.

Key words: Jackknife; Resampling; Sampling Variability; Inequality

JEL: C81, C87, D3

♠ Christian Albrechts University of Kiel, Department of Economics. Phone: +49 (0)431 8801772. Email: carsten.schroeder@bwl.uni-kiel.de. We would like to thank Stephen Jenkins for his most helpful comments. Further, we would like to thank Friedrich Bergmann and Jan Krause for research assistance.

1 Introduction

When examining time-series changes in inequality or cross country differences in inequality, the measured changes are sometimes small. To estimate the precision of a statistic from a sample and to test the statistical significance of changes or cross country differences of the same statistic, the jackknife has been suggested.[1] The jackknife is a resampling method that uses subsets of the original database by leaving out one observation at a time from the sample. So, there are as many subsets as there are observations in the sample, and for each subset the jackknife statistic needs to be computed. This means that the jackknife can become a time intensive procedure when the sample size is large.

Figure 1 provides an illustration of this problem for a hypothetical income distribution. Particularly, it shows the computer time needed for deriving several well-known inequality indices[2] from all the jackknife subsets using the frequently used STATA software 'inequal7.ado.'[3] We start with a sample size of four, which then is always doubled. For small sample sizes (<5000) it takes only a few second to compute all the jackknife indices. For larger sample sizes, however, computer time increases exponentially: for 10,240 observations jackknifing takes 2.52 minutes; for 20,480 observations it takes 9.82 minutes; and for 81,920 it takes 261.70 minutes. Many comparative inequality analyses step on data from several points in time, countries and income concepts. Then implementing the jackknife can take days or weeks. This is a serious problem for researchers who need to access data that are stored on external servers, and who face limited processing power for their computations as defined by the data provider. The well-known Luxembourg Income Study is one important example.

Figure 1 about here

[1] For the theoretical justification for the jackknife and other related resampling techniques see Efron (1982).
[2] The inequality indices are: relative mean deviation; coefficient of variation; standard deviation of logs; Gini coefficient; Mehran index; Piesch index; Kakwani index; Theil index; Mean Log Deviation; Generalized entropy measures with sensitivity parameter -1 and 2.
[3] The software is run on the following hardware: 64-bit system; 8 GB ram; core(TM)2 Duo CPU; 3GHz.

Karagiannis and Kovacevic (2000) and Yitzhaki (1991), however, show that jackknifing the Gini coefficient requires only a few passes through the data.[4] We complement these two works on the Gini coefficient by providing efficient jackknife procedures for several frequently applied inequality indices: the coefficient of variation, the variance of the logarithms, the mean log deviation, the Theil index, and the Atkinson index.[5] We show that, after having computed some basic statistics from the overall sample, it takes a single run through the data to derive all the jackknife values of an index from all the subsets. We apply the outlined jackknife procedure to micro data from the Luxembourg Income Study.

Section 2 explains the procedures. Section 3 provides the results from the empirical application. Section 4 concludes. Derivations of all the formulas and STATA codes are provided in an Appendix.

2 Efficient jackknife procedures for inequality indices

The jackknife offers a conceptually simple way to estimate the precision of a statistic (see the pioneering works of Tukey, 1958; Efron, 1982; Efron and Gong, 1983; Wolter, 1985). In the context of inequality measurement, we have a random sample of N observations on income, $\boldsymbol{y} = (y_1, y_2, \dots, y_N)$ and sampling weights, $\omega_1, \omega_2, \dots, \omega_N$. Let $\theta = \theta(\boldsymbol{y})$ denote our measure of inequality. Let $\theta_{(i)} = \theta(y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_N)$ denote the jackknife estimate of the same measure of inequality for the subset where the ith observation has been deleted.

Following Wolter (1985), the jackknife estimate of the standard error of θ is,

$$(1) \quad SE_\theta = \left(\frac{N-1}{N} \sum_{i=1}^{N} \frac{\omega_i}{\bar{\omega}} \left[\theta_{(i)} - \theta \right]^2 \right)^{0.5},$$

[4] Ogwang (2000) shows that it is also possible to obtain standard errors for the Gini index from OLS regression. Giles (2004) extends the regression-based approach to test hypotheses regarding the sensitivity of the Gini coefficient to changes in the data using seemingly unrelated regressions.

[5] Karoly (1989) also derives jackknife procedures for calculating the between- and within-group inequality components of the variance of the logarithms, the mean log deviation, and the Theil index.

with $\bar{\omega} = \frac{1}{N}\sum_{i=1}^{N} \omega_i$.[6] Computing the jackknife standard error estimate relies on the N values of $\theta_{(i)}$, one jackknife statistic per subset. For large samples the computational burden to derive equation (1) seems to be large. However, as we will outline below, for standard inequality indices deriving the N values of $\theta_{(i)}$ requires just a few passes through the data. Hereby, the number of passes is independent of the number of sample observations, N.

The procedure is detailed below by means of the Theil index, and the variance of logarithms. The general idea of the procedure is to write the jackknife estimates $\theta_{(1),\dots,}\theta_{(N)}$ as a function of statistics from the overall sample (i.e., as a function of θ, N, arithmetic or geometric mean) and a subset-specific correction factor that can be derived with a single run through the data. The procedure can be adapted to other inequality indices including indices of the generalized entropy class, and indices based on the variance or social-welfare functions (e.g. the Atkinson index).

We will make use of the following notation and definitions:

1. w_i denotes the normalized weight, $w_i = \frac{\omega_i}{\frac{1}{N}\sum_{i=1}^{N} \omega_i}$. Accordingly, $\sum_{i=1}^{N} w_i = N$.
2. $\bar{y}$ denotes the arithmetic mean of income, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} w_i y_i$.
3. y^* denotes the geometric mean of income, $y^* = \exp\left(\frac{1}{N}\sum_{i=1}^{N} w_i \ln(y_i)\right)$. The natural logarithm of the geometric mean is denoted $\bar{x} = \ln(y^*) = \frac{1}{N}\sum_{i=1}^{N} w_i x_i$ with $x_i = \ln(y_i)$.

[6] An alternative method is to compute the squared differences between the jackknife statistics and their mean (see, for example, Yitzhaki, 1991).

2.1 Efficient jackknife procedure for the Theil index

The Theil index from the sample is,

$$(2) \quad \theta_T = \frac{1}{N\bar{y}}\left(\sum_{i=1}^{N} w_i y_i \ln(y_i)\right) - \ln(\bar{y}).$$

The Theil index for the subset where the ith observation has been deleted is,

$$(3) \quad \theta_{T(i)} = \frac{1}{(N - w_i)\bar{y}_{(i)}}\left(\sum_{j \neq i} w_j y_j \ln(y_j)\right) - \ln(\bar{y}_{(i)}),$$

with $\bar{y}_{(i)}$ denoting the arithmetic mean of income from the subset,

$$(4) \quad \bar{y}_{(i)} = \frac{N\bar{y} - w_i y_i}{N - w_i}.$$

The first step is to write $\theta_{T(i)}$ in terms of θ_T. Initially, from (3):

$$(5) \quad \theta_{T(i)} = \frac{1}{(N - w_i)\bar{y}_{(i)}}\left(\sum_{i=1}^{N} w_i y_i \ln(y_i)\right) - \frac{w_i}{(N - w_i)}\frac{y_i}{\bar{y}_{(i)}} ln(y_i) - \ln(\bar{y}_{(i)}).$$

Rewriting equation (2) gives,

$$(6) \quad \sum_{i=1}^{N} w_i y_i \ln(y_i) = [\theta_T + \ln(\bar{y})]N\bar{y},$$

and substituting (6) and (4) into (5) gives,

$$(7) \quad \theta_{T(i)} = \frac{N\bar{y}}{N\bar{y} - w_i y_i}\left(\theta_T + \ln(\bar{y})\right) - \frac{w_i y_i \ln(y_i)}{N\bar{y} - w_i y_i} - ln\left(\frac{N\bar{y} - w_i y_i}{N - w_i}\right).$$

Equation (7) reveals that $\theta_{T(i)}$ can be expressed as a function of three statistics from the full sample, $N, \bar{y},$ and θ_T, and characteristics of the observation that is left out, w_i and y_i. Thus, after having calculated $N, \bar{y},$ and θ_T for the full sample, to compute all the jackknife statistics $\theta_{T(1)}, \ldots, \theta_{T(N)}$ takes a single pass through the data.

2.2 Efficient jackknife procedure for the variance of logarithms

Applying Bessel's correction[7], the variance of the logarithms from the sample is,

$$(8)\quad \theta_{VL} = \frac{1}{N-1}\sum_{i=1}^{N} w_i \ln\left(\frac{y_i}{y^*}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N} w_i (x_i - \bar{x})^2$$

The variance of the logarithms for the subset where the ith observation has been deleted is,

$$(9)\quad \theta_{VL(i)} = \frac{1}{N-2}\sum_{j\neq i} w_{j(i)} (x_j - \bar{x}_{(i)})^2,$$

with $\bar{x}_{(i)} = \frac{1}{N-w_i}[N\bar{X} - x_i w_i]$, and with $w_{j(i)} = \frac{w_j}{(N-w_i)/(N-1)}$ denoting re-weighted normalized weights. By means of the re-weighting the average of $w_{j(i)}$ over the subset where the ith observation has been deleted equals unity. So, the analogue of the term $\frac{1}{N-1}$ in (8) in (9) is $\frac{1}{N-2}$. Substituting the definition of $w_{j(i)}$ in (9) gives:

$$(10)\quad \theta_{VL(i)} = \frac{(N-1)}{(N-2)(N-w_i)}\sum_{j\neq i} w_j (x_j - \bar{x}_{(i)})^2 ,$$

Initially, from (8):

$$(11)\quad \theta_{VL} = \frac{1}{N-1}\sum_{j\neq i}^{N} w_j (x_j - \bar{x})^2 + \frac{1}{N-1} w_i (x_i - \bar{x})^2.$$

Substituting $\bar{x} = \frac{1}{N}\left[(N-w_i)\bar{x}_{(i)} + x_i w_i\right]$ in (11) gives:

$$(12)\quad \theta_{VL} = \frac{1}{N-1}\sum_{j\neq i}^{N} w_j \left(x_j - \frac{1}{N}\left[(N-w_i)\bar{x}_{(i)} + x_i w_i\right]\right)^2 + \frac{1}{N-1} w_i (x_i - \bar{x})^2$$

$$= \frac{1}{N-1}\sum_{j\neq i}^{N} w_j \left(\underbrace{x_j - \frac{N}{N}\bar{x}_{(i)}}_{\boldsymbol{A}} + \underbrace{\frac{w_i}{N}\bar{x}_{(i)} - \frac{w_i}{N}x_i}_{\boldsymbol{B}}\right)^2 + \frac{1}{N-1} w_i (x_i - \bar{x})^2$$

Equation (12) can be rewritten as:

[7] Bessel's correction, the division in the variance formula by $N-1$ instead of by N, secures unbiasdness.

$$(13)\quad \theta_{VL} = \underbrace{\frac{1}{N-1}\sum_{j\neq i}^{N} w_j\left(x_j - \bar{x}_{(i)}\right)^2}_{\boldsymbol{C}} + \underbrace{\frac{2}{(N-1)}\sum_{j\neq i}^{N} w_j\left(\left(x_j - \bar{x}_{(i)}\right)\left(\frac{w_i}{N}\right)\left(\bar{x}_{(i)} - x_i\right)\right)}_{\boldsymbol{D}}$$

$$+ \underbrace{\frac{1}{N-1}\sum_{j\neq i}^{N} w_j\left(\frac{w_i}{N}\bar{x}_{(i)_j} - \frac{w_i}{N}x_i\right)^2}_{\boldsymbol{E}} + \frac{1}{N-1}w_i(x_i - \bar{x})^2$$

The C -term on the right handside of (12) can be rewritten as $= \theta_{VL(i)}\frac{(N-2)(N-w_i)}{(N-1)^2}$. The D -term is zero since

$$(14)\quad D = \frac{2}{N-1}\frac{w_i}{N}\sum_{j\neq i} w_j\left(x_j - \bar{x}_{(i)}\right)\left(\bar{x}_{(i)} - x_i\right) = \frac{2\,w_i}{(N-1)N}\left(\bar{x}_{(i)} - x_i\right)\underbrace{\sum_{j\neq i} w_j\left(x_j - \bar{x}_{(i)}\right)}_{=0} = 0$$

The E –term after some algebra becomes,

$$(15)\quad E = \frac{1}{N-1}\frac{{w_i}^2}{(N-w_i)^2}\sum_{j\neq i} w_j(\bar{x} - x_i)^2$$

$$= \frac{1}{N-1}\frac{{w_i}^2}{(N-w_i)^2}(N-w_i)(\bar{x} - x_i)^2 = \frac{1}{N-1}\frac{{w_i}^2}{N-w_i}(\bar{x} - x_i)^2$$

Substituting (14-15) in (13), the variance of the logarithms for the sample becomes,

$$(16)\quad \theta_{VL} = \theta_{VL(i)}\frac{(N-2)(N-w_i)}{(N-1)^2} + \frac{1}{N-1}\frac{{w_i}^2}{N-w_i}(\bar{x} - x_i)^2 + \frac{w_i}{N-1}(x_i - \bar{x})^2.$$

After some algebra, (16) becomes,

$$(17)\quad \theta_{VL} = \theta_{VL(i)}\frac{(N-2)(N-w_i)}{(N-1)^2} + \frac{Nw_i}{(N-1)(N-w_i)}(\bar{x} - x_i)^2.$$

Solving (17) with respect to $\theta_{VL(i)}$ gives the desired expression for the jackknife estimator of the variance of the logarithms:

$$(18)\quad \theta_{VL(i)} = \theta_{VL}\frac{(N-1)^2}{(N-2)(N-w_i)} - \frac{Nw_i(N-1)}{(N-w_i)^2(N-2)}(\bar{x} - x_i)^2$$

Equation (18) is the analogue of the jackknife estimator of the Theil index in equation (7): $\theta_{VL(i)}$ can be expressed as a function of statistics from the full sample ($N, \bar{x},$ and θ_{VL}) and the characteristics of the observation that is left out, w_i and x_i. Thus, after having calculated $N, \bar{x},$ and θ_{VL} for the full sample, computing $\theta_{VL(1)}, \dots, \theta_{VL(N)}$ takes a single pass through the data.

2.3 Efficient jackknife procedure for other inequality indices

Similar derivations as those explained in Sections 2.1 and 2.2 can be made for other inequality indices. Formulas for an efficient computation of the Atkinson index, θ_{A_ε} (with inequality aversion parameter $\varepsilon = 1$ and $\varepsilon = 2$), the mean log deviation, θ_{MLD}, and the coefficient of variation, θ_{CV}, are as follow:

$$(19) \quad \theta_{A_1(i)} = 1 - \frac{exp\left[\frac{N}{N-w_i}\ln(y^*) - \frac{\ln(y_i)\,w_i}{N-w_i}\right]}{\frac{N\bar{y}-w_i y_i}{N-w_i}}$$

$$(20) \quad \theta_{A_2(i)} = 1 - \frac{N-w_i}{\frac{N\bar{y}-w_i y_i}{\bar{y}(N-w_i)}\frac{N}{1-\theta_{A_2}} - \frac{w_i(N\bar{y}-w_i y_i)}{y_i(N-w_i)}}$$

$$(21) \quad \theta_{MLD(i)} = \frac{1}{N-w_i}[\theta_{MLD} - \ln(\bar{y})] + \frac{w_i \ln(y_i)}{N-w_i} + \ln\left(\frac{N\bar{y}-y_i w_i}{N-w_i}\right)$$

$$(22) \quad \theta_{CV(i)} = \frac{\left(\theta_V \frac{(N-1)^2}{(N-2)(N-w_i)} - \frac{N w_i (N-1)}{(N-w_i)^2(N-2)}(\bar{y}-y_i)^2\right)^{0.5}}{\frac{1}{(N-w_i)}[(N\bar{y}-y_i w_i)]}$$

Derivations of the formulas can be found in the Appendix. Again, after having calculated some basic statistics from the full sample, computing all the jackknife indices takes only a single pass through the data.

3 Empirical application

We have calculated the above inequality indices and their associated jackknife confidence intervals for distributions of disposable household incomes in the US and in Germany from the Luxembourg Income Study (LIS) database. For 40 countries and several years, the LIS provides representative micro-level information on private households' incomes and their demographics.

Our computations rely on the LIS household-level datasets. Household disposable income is our income concept. Household disposable income is harmonized across countries, covers labor earnings, property income, and government transfers in cash minus income and payroll taxes. To adjust household incomes for differences in needs, we have deflated household disposable income by means of the square root equivalence scale. The square root equivalence scale is the number of household members to the power of 0.5. This gives the needs-adjusted equivalent income of the household. Household units are weighted by the frequency weights (as provided in the data) and the number of household members. Our weighting procedure accommodates the principle of normative individualism that considers any person as important as any other. The so derived distribution depicts differences in living standards, captured by differences in equivalent incomes, among individuals (Bönke and Schröder, 2012).

We have removed household observations with missing information or with negative values of disposable income. Moreover, to avoid outlier-driven biases of inequality estimates, we use trimmed data with the one percent observations with the highest and with the lowest incomes being discarded.

It has taken a few seconds to obtain all the results presented in Table 1. The Table is split in two panels. The upper panel provides the results for the US, the lower panel provides the results for Germany. In the US, the results cover the period 1991-2010; in Germany, the results cover the period 1994-2010. For every country-period combination, the Table provides the point estimates of the inequality indices along with their upper and lower bounds of 95 percent confidence intervals, CI_{θ}^{lo} and CI_{θ}^{hi}, derived from the jackknife statistics.

Table 1 about here

We comment on the US first. An examination of the statistics shows a significant rise of inequality over the observation period: the point estimate of the Theil index increases from 0.161 in 1991 to 0.192 in 2010, and the confidence intervals are clearly distinct: $[0.158; 0.165]$ vs. $[0.189; 0.196]$. However, some inter-temporal changes in inequality for this sample are not statistically significant (e.g. 1997-2000; 2000-2004; 2004-2007).

For Germany, we also see a significant rise of inequality over the observation period. This is due to a prominent rise of inequality between 2000 and 2004. The inter-temporal comparisons before the rise (1994-2000) and after the rise (2004-2007 and 2007-2010) indicate no significant changes in inequality.

Comparing inequality levels in the US and Germany there is significantly more inequality in the US. The result holds for all six inequality indices and all the observed points in time.[8]

4 Conclusion

This paper has outlined a procedure to obtain jackknife estimates for several inequality indices with only a few passes through the data. The number of passes is independent of the number of observations: After having computed some statistics from the overall sample, computing all the jackknife indices takes only a single pass through the data. Hence, the method provides an efficient way to get standard errors of the estimators even if sample size is large.

We have applied our method using data from the Luxembourg Income Study to evaluate the statistical significance of inter-temporal inequality in Germany and the US, and also to evaluate cross country differences in inequality levels.

[8] We have executed our empirical analysis using the alternative formulation of the standard error introduced in footnote 4. It did not change our conclusions since confidence intervals changed very tittle.

References

Schröder, C., and T. Bönke (2012). Country Inequality Rankings and Conversion Schemes. Economics: The Open-Access, Open-Assessment E-Journal, Vol. 6, 2012-28.

Efron, B. (1982): The Jackknife, the Bootstrap and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia PA.

Efron, B., and G. Gong (1983): A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation, The American Statistician, 37, 36-48.

Giles, D. (2004): Calculating a Standard Error for the Gini Coefficient: Some Further Results, Oxford Bulletin of Economics and Statistics, 66, 425-433.

Karagiannis, E., and M. Kovacevic (2000): Practitioners Corner - A Method to Calculate the Jackknife Variance Estimator for the Gini Coefficient, Oxford Bulletin of Economics and Statistics, 62, 199-122.

Karoly, L.A. (1988): Computing Standard Errors for Measures of Inequality using the Jackknife, unpublished manuscript.

Karoly, L.A. (1992): Changes in the Distribution of Individual Earnings in the United States: 1967-1986, The Review of Economics and Statistics, 74, 107-115.

Luxembourg Income Study (LIS) Database, http://www.lisdatacenter.org (Germany and US; 1991-2010). Luxembourg: LIS.

Ogwang, T. (2000): A Convenient Method of Computing the Gini Index and its Standard Error, Oxford Bulletin of Economics and Statistics, 62, 123-29.

Tukey, J. W. (1958): Bias and confidence in not quite large samples. Annals of Mathematical Statisics, 29, 614.

Wolter, K. (1985): Introduction to Variance Estimation, Springer, New York.

Yitzhaki, S. (1991): Calculating Jackknife Variance Estimators for Parameters of the Gini Method, Journal of Business and Economic Statistics, 9, 235-239.

Appendix

A.1 Derivation of jackknife formulas

Mean log deviation (Entropy 0)

$$(1^{MLD}) \quad \theta_{MLD} = \frac{1}{N}\sum_{i=1}^{N} w_i \ln\left(\frac{\bar{y}}{y_i}\right) = -\frac{1}{N}\sum_{i=1}^{N} w_i \ln(y_i) + \ln(\bar{y})$$

$$(2^{MLD}) \quad \theta_{MLD(i)} = -\frac{1}{N - w_i}\sum_{j\neq i} w_j \ln(y_i) + \ln(\bar{y})$$

From (2^{MLD}):

$$(3^{MLD}) \quad \theta_{MLD(i)} = -\frac{1}{N - w_i}\left[\sum_{j\neq i} w_j \ln(y_j) + w_i \ln(y_i)\right] + \frac{w_i \ln(y_i)}{N - w_i} + \ln(\bar{y}_{(i)})$$

$$(4^{MLD}) \quad \theta_{MLD(i)} = -\frac{1}{N - w_i}\left[\sum_{i=1}^{N} w_i \ln(y_j)\right] + \frac{w_i \ln(y_i)}{N - w_i} + \ln(\bar{y}_{(i)})$$

Substituting $-N[\theta_{MLD} - \ln(\bar{y})] = \sum_{i=1}^{N} w_i \ln(y_i)$ from (1^{MLD}) gives:

$$(5^{MLD}) \quad \theta_{MLD(i)} = -\frac{1}{N - w_i}\left[-N[\theta_{MLD} - \ln(\bar{y})]\right] + \frac{w_i \ln(y_i)}{(N - w_i)} + \ln(\bar{y}_{(i)})$$

Substituting $\bar{y}_{(i)}$ by $\frac{N\bar{y} - y_i w_i}{N - w_i}$ gives:

$$(6^{MLD}) \quad \theta_{MLD(i)} = \frac{1}{N - w_i}[\theta_{MLD} - \ln(\bar{y})] + \frac{w_i \ln(y_i)}{N - w_i} + \ln\left(\frac{N\bar{y} - y_i w_i}{N - w_i}\right)$$

Atkinson Index

The general form of the Atkinson index is, $\theta_{A_\varepsilon} = 1 - \left[\frac{1}{N}\sum_{i=1}^{N} w_i \left(\frac{\bar{y}}{y_i}\right)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}}$. Below we derive the jackknife formulas for two prominent case of the inequality aversion parameter, ε.

Inequality aversion parameter $\boldsymbol{\varepsilon = 1}$

$$(1^{A_1}) \quad \theta_{A_1} = 1 - \frac{y^*}{\bar{y}} \quad = \quad 1 - \frac{\exp\left[\frac{1}{N}\sum_{i=1}^{N} w_i \ln(y_i)\right]}{\bar{y}}$$

$$(2^{A_1}) \quad \theta_{A_1(i)} = 1 - \frac{\exp\left[\frac{1}{N - w_i}\sum_{j\neq i} w_j \ln(y_j)\right]}{\bar{y}_{(i)}}$$

Expansion of the term in brackets in the numerator with $\frac{\ln(y_i)w_i}{N-w_i} - \frac{\ln(y_i)w_i}{N-w_i}$, and substitution of $\bar{y}_{(i)}$ by $\frac{N\bar{y}-y_i w_i}{N-w_i}$ gives:

$$(3^{A_1}) \quad \theta_{A_1(i)} = 1 - \frac{exp\left[\frac{N}{N - w_i}\left(\frac{1}{N}\sum_{i=1}^{N} w_i \ln(y_i)\right) - \frac{\ln(y_i)\, w_i}{N - w_i}\right]}{\frac{N\bar{y} - w_i y_i}{N - w_i}}$$

Substitution of the term $\frac{1}{N}\sum_{i=1}^{N} w_i \ln(y_i)$ (log of the geometric mean of income from the full sample) by $\ln(y^*)$ gives:

$$(4^{A_1}) \quad \theta_{A_1(i)} = 1 - \frac{exp\left[\frac{N}{N - w_i}\ln(y^*) - \frac{\ln(y_i)\, w_i}{N - w_i}\right]}{\frac{N\bar{y} - w_i y_i}{N - w_i}}$$

Inequality aversion parameter $\boldsymbol{\varepsilon = 2}$

$$(1^{A_2}) \quad \theta_{A_2} = 1 - \frac{N}{\sum_{i=1}^{N} w_i \frac{\bar{y}}{y_i}}$$

$$(2^{A_2}) \quad \theta_{A_2(i)} = 1 - \frac{N - w_i}{\sum_{j\neq i}^{N} w_j \frac{\bar{y}_{(i)}}{y_j}}$$

Expansion of the denominator with $w_i \frac{\bar{y}_{(i)}}{\bar{y}} \frac{\bar{y}}{y_i} - w_i \frac{\bar{y}_{(i)}}{\bar{y}} \frac{\bar{y}}{y_i}$ and rewriting the sum gives:

$$(3^{A_2}) \quad \theta_{A_2(i)} = 1 - \frac{N - w_i}{\left(\frac{\bar{y}_{(i)}}{\bar{y}} \sum_{j \neq i} w_j \frac{\bar{y}}{y_j}\right) + w_i \frac{\bar{y}_{(i)}}{\bar{y}} \frac{\bar{y}}{y_i} - w_i \frac{\bar{y}_{(i)}}{\bar{y}} \frac{\bar{y}}{y_i}}$$

$$(4^{A_2}) \quad \theta_{A_2(i)} = 1 - \frac{N - w_i}{\left(\frac{\bar{y}_{(i)}}{\bar{y}} \sum_{i=1}^{N} w_i \frac{\bar{y}}{y_i}\right) - w_i \frac{\bar{y}_{(i)}}{\bar{y}} \frac{\bar{y}}{y_i}}$$

From $\theta_{A_2} = 1 - \frac{N}{\sum_{i=1}^{N} w_i \frac{\bar{y}}{y_i}}$ it follows that $\sum_{i=1}^{N} \frac{w_i \bar{y}}{y_i} = \frac{N}{1 - \theta_{A_2}}$, and replacement of the sum in the denominator gives:

$$(5^{A_2}) \quad \theta_{A_2(i)} = \quad 1 - \quad \frac{N - w_i}{\frac{\bar{y}_{(i)}}{\bar{y}} \frac{N}{1 - \theta_{A_2}} - \left(\frac{w_i \bar{y}_{(i)}}{y_i}\right)}$$

Finally, substitution of $\bar{y}_{(i)}$ by $\frac{N\bar{y} - y_i w_i}{N - w_i}$ gives:

$$(6^{A_2}) \quad \theta_{A_2(i)} = \quad 1 - \frac{N - w_i}{\frac{N\bar{y} - w_i y_i}{\bar{y}(N - w_i)} \frac{N}{1 - \theta_{A_2}} - \frac{w_i (N\bar{y} - w_i y_i)}{y_i (N - w_i)}}$$

Variance and Coefficient of Variation

$$(1^V) \quad \theta_V = \frac{1}{N-1}\sum_{i=1}^{N} w_i(y_i - \bar{y})^2$$

$$(2^V) \quad \theta_{V(i)} = \frac{(N-1)}{(N-2)(N-w_i)}\sum_{j\neq i} w_j(y_j - \bar{y}_{(i)})^2$$

Rewriting of θ_V gives:

$$(3^V) \quad \theta_V = \frac{1}{N-1}\sum_{j\neq i} w_j(y_j - \bar{y}_{(i)})^2 + \frac{1}{N-1} w_i[y_i - \bar{y}]^2$$

Substituting $\bar{y} = \frac{1}{N}\left[(N-w_i)\bar{y}_{(i)} + y_i w_i\right]$ and reorganizing in analogy to the variance of the logarithms gives:

$$(4^V) \quad \theta_V = C + D + E + \frac{1}{N-1} w_i(y_i - \bar{y})^2$$

$$(5^V) \quad C = \frac{1}{N-1}\sum_{j\neq i} w_j(y_j - \bar{y}_{(i)})^2 = \theta_{V(i)}\frac{(N-2)(N-w_i)}{(N-1)^2}$$

$$(6^V) \quad D = \frac{2}{N-1}\frac{w_i}{N}\sum_{j\neq i} w_j(y_j - \bar{y}_{(i)})(\bar{y}_{(i)} - y_i)$$

$$= \frac{2\,w_i}{(N-1)N}(\bar{y}_{(i)} - y_i)\underbrace{\sum_{j\neq i} w_j(y_j - \bar{y}_{(i)})}_{=0} = 0$$

$$(7^V) \quad E = \frac{1}{N-1}\left(\frac{w_i}{N}\right)^2\sum_{j\neq i} w_j(\bar{y}_{(i)} - y_i)^2$$

Analogously to θ_{VL}we can rewrite (7^V) as:

$$(8^V) \quad E = \frac{1}{N-1}\frac{{w_i}^2}{N-w_i}(\bar{y} - y_i)^2$$

Substituting (5^V), (6^V), and (8^V) in (4^V) gives:

$$(9^V) \quad \theta_V = \theta_{V(i)} \frac{(N-2)(N-w_i)}{(N-1)^2} + \frac{1}{N-1} \frac{{w_i}^2}{N-w_i} (\bar{y} - y_i)^2 + \frac{w_i}{N-1} (y_i - \bar{y})^2$$

Analogously to θ_{VL}we can rewrite (9^V) as:

$$(9^V) \quad \theta_V = \theta_{V(i)} \frac{(N-2)(N-w_i)}{(N-1)^2} + \frac{Nw_i}{(N-1)(N-w_i)} (\bar{y} - y_i)^2$$

Solving (9^V) for $\theta_{V(i)}$ gives:

$$(10^V) \quad \theta_{V(i)} = \theta_V \frac{(N-1)^2}{(N-2)(N-w_i)} - \frac{Nw_i(N-1)}{(N-w_i)^2(N-2)} (\bar{y} - y_i)^2$$

The coefficient of variation is defined as,

$$(1^{CV}) \quad \theta_{CV} = \frac{(\theta_V)^{0.5}}{\bar{y}}$$

Hence,

$$(2^{CV}) \quad \theta_{CV(i)} = \frac{\left(\theta_{V(i)}\right)^{0.5}}{\bar{y}_{(i)}}$$

Substitution of $\theta_{V(i)} = \theta_V \frac{(N-1)^2}{(N-2)(N-w_i)} - \frac{Nw_i(N-1)}{(N-w_i)^2(N-2)} (\bar{y} - y_i)^2$ and of $\bar{y}_{(i)} = \frac{1}{(N-w_i)} [(N\bar{y} - y_i w_i)]$ in (2^{CV}) gives,

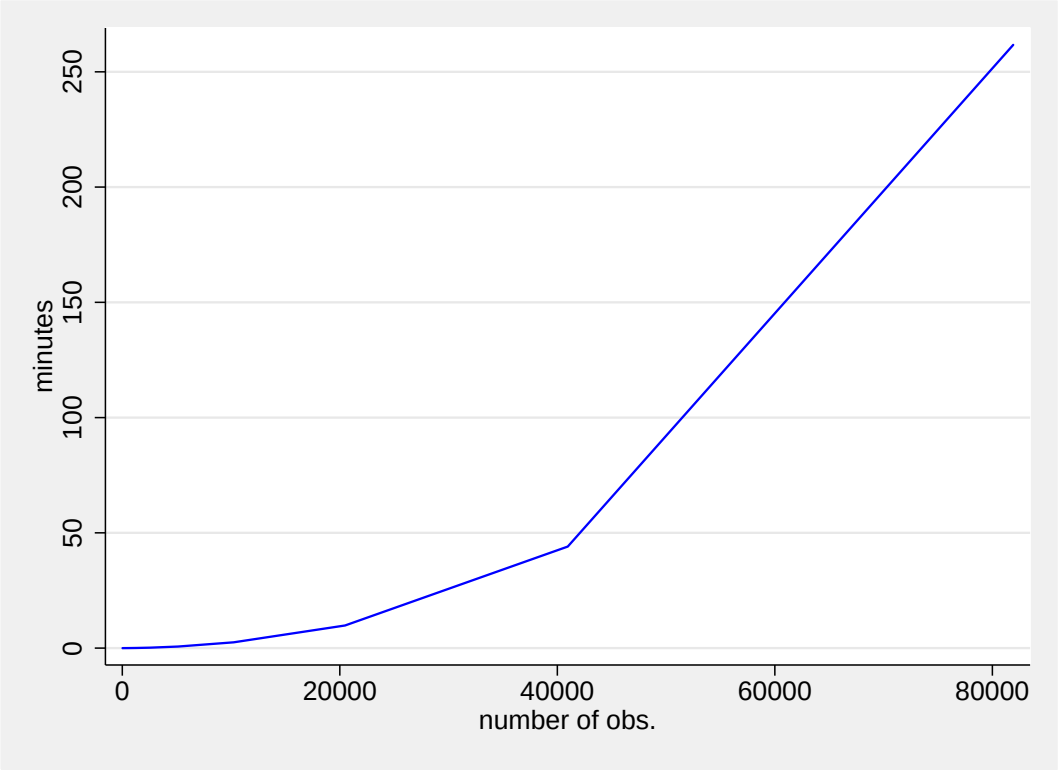
$$(3^{CV}) \quad \theta_{CV(i)} = \frac{\left(\theta_V \frac{(N-1)^2}{(N-2)(N-w_i)} - \frac{Nw_i(N-1)}{(N-w_i)^2(N-2)} (\bar{y} - y_i)^2 \right)^{0.5}}{\frac{1}{(N-w_i)} [(N\bar{y} - y_i w_i)]}$$

A.2 STATA code for Luxembourg Income Study

```
#delimit ;

*** loop over countries;
foreach file in $us91h $us97h $us00h $us04h $us07h $us10h $de94h $de00h $de04h $de07h $de10h {;
        * Variables of interest;
        local vars "dname did hwgt dhi nhhmem";
        * open data;
        use `vars' using `file', clear;
        ********************************************;
        * Data preparation and auxiliary statistics *;
        ********************************************;
        qui rename hwgt w;
        qui rename dhi y;
        * drop negative or zero yomes (because of log);
        qui drop if y==. | y<=0;
        * trimming top bottom 1percent of unweighted observations;
        xtile centiles=y, nq(100);
        drop if centiles ==1 | centiles==100;
        * drop missings;
        qui drop if nhhmem==. | w==.;
        * weight by frequency weights and number of household members;
        qui replace w=w*nhhmem;
        * compute equivalent yome using square root scale;
        qui replace y=y/(nhhmem)^(0.5);
        qui gen logy=log(y);
        * Normalization of the weights;
        qui sum w;
        qui replace w=w/r(mean);
        * Arithmetic mean (weighted);
        qui sum y [w=w];
        qui scalar sc_mu=r(mean);
        * geometric mean yome (weighted);
        qui gen help=logy*w;
        qui sum help;
        qui scalar sc_gmu=exp(r(mean));
        qui drop help;
        * Sample size (weighted);
        qui scalar sc_N=r(N);
        ********************************************;
        *** Inequality indices from overall sample **;
        ********************************************;
        *Atkinson Index 1: stored in scalar sc_A1 ***;
        qui gen summand=w*ln(y);
        qui sum summand;
        qui scalar sc_gmu=exp(r(sum)/sc_N);
        qui scalar sc_A1=1-sc_gmu/sc_mu;
        qui drop summand;
        *Atkinson Index 2: stored in scalar sc_A2 ***;
        qui gen summand=w*(y/sc_mu)^(1-2);
        qui sum summand;
        qui scalar sc_A2=1-(r(sum)/sc_N)^(1/(1-2));
        qui drop summand;
        *Mean log deviation: stored in scalar sc_MLD*;
        qui gen summand=w*ln(y);
        qui sum summand;
        qui scalar sc_MLD=-r(sum)/(sc_N)+ln(sc_mu);
        qui drop summand;
        *Theil index: stored in scalar sc_T*;
```

```
		qui gen summand=y/sc_mu*ln(y)*w;
		qui sum summand;
		qui scalar sc_T=r(mean)-ln(sc_mu);
		qui drop summand;
		*Variance of log yomes: stored in scalar sc_V*;
		qui gen summands=(logy-log(sc_gmu))^2*w;
		qui sum summands;
		qui scalar sc_VL=r(sum)/(sc_N-1);
		qui drop summands;
		*Variance and coeff of var: stored in scalar sc_V and sc_CV*;
		qui gen summands=(y-sc_mu)^2*w;
		qui sum summands;
		qui scalar sc_V=[r(sum)/(sc_N-1)];
		qui scalar sc_CV=sc_V^(0.5)/sc_mu;
		qui drop summands;

		*********************************************;
		**** Inequality indices from JK samples *****;
		*********************************************;
		*Atkinson Index 1: stored in variable jk_A1 ***;
		qui gen jk_A1=1-exp(sc_N/(sc_N-w)*ln(sc_gmu)-ln(y)*w/(sc_N-w))/((sc_N*sc_mu-w*y)/(sc_N-w));
		*Atkinson Index 2: stored in variable jk_A2 ***;
		qui gen jk_A2=1-(sc_N-w)/[(sc_N*sc_mu-w*y)/(sc_mu*(sc_N-w))*sc_N/(1-sc_A2)-w*(sc_N*sc_mu-
w*y)/(y*(sc_N-w))];
		*Mean log deviation: stored in variable jk_MLD ***;
		qui gen jk_MLD=sc_N/((sc_N-w))*(sc_MLD-ln(sc_mu))+w*ln(y)/(sc_N-w)+ln((sc_N*sc_mu-y*w)/(sc_N-w));
		*Theil index: stored in variable jk_T ***;
		qui gen jk_T=(sc_N*sc_mu)/((sc_N*sc_mu-w*y))*(sc_T+ln(sc_mu))-(w*y*ln(y))/((sc_N*sc_mu-w*y))-
ln((sc_N*sc_mu-w*y)/(sc_N-w));
		*Variance of logs: stored in variable jk_VL ***;
		qui gen jk_VL=(sc_N-1)^2/((sc_N-2)*(sc_N-w))*sc_VL-sc_N*w*(sc_N-1)/((sc_N-w)^2*(sc_N-2))*(log(sc_gmu)-
logy)^2;
		*Variance: stored in variable jk_V ***;
		qui gen jk_V=(sc_N-1)^2/((sc_N-2)*(sc_N-w))*sc_V-sc_N*w*(sc_N-1)/[(sc_N-2)*(sc_N-w)^2]*(y-sc_mu)^2;
		*Coefficient of var: stored in variable jk_V ***;
		qui gen jk_CV=(jk_V)^(0.5)/((sc_N*sc_mu-y*w)/(sc_N-w));

		********** 95% normal confidence interval **********;
		**** using normalized weights as in WOLTER (1985) to compute variance;
		local vars "A1 A2 MLD T VL CV";
		* loop over inequality indices;
		foreach var of local vars {;
			qui gen jk_V_`var'=((sc_N-1)/(sc_N)*w*(sc_`var'-jk_`var')^2);
			qui sum jk_V_`var';
			qui scalar sc_V_`var'=r(sum);
			qui scalar sc_SD_`var'=sc_V_`var'^(0.5);
			qui scalar sc_lo_`var' =sc_`var'-1.96*sc_SD_`var';
			qui scalar sc_hi_`var' =sc_`var'+1.96*sc_SD_`var';
			disp dname " `var' " " lower_bound " sc_lo_`var' " Point estimate " sc_`var' " upper_bound " sc_hi_`var' ;
		};
};
******;
```

Figure 1. Computer time and sample size



Note. Own computations. The jackknife has been implemented using STATA's software package inequal7.ado on a computer with characteristics: 64-bit system; 8 GB ram; core(TM)2 Duo CPU; 3GHz. See also footnotes 2 and 3.

Table 1. Inequality indices

		Atkinson $\epsilon = 1$			Atkinson $\epsilon = 2$			Mean log deviation			Theil index			Variance of logs			Coeff. of variation		
	Year	$CI^{lo}_{\theta_{A_1}}$	θ_{A_1}	$CI^{hi}_{\theta_{A_1}}$	$CI^{lo}_{\theta_{A_2}}$	θ_{A_2}	$CI^{hi}_{\theta_{A_2}}$	$CI^{lo}_{\theta_{MLD}}$	θ_{MLD}	$CI^{hi}_{\theta_{MLD}}$	$CI^{lo}_{\theta_T}$	θ_T	$CI^{hi}_{\theta_T}$	$CI^{lo}_{\theta_{VL}}$	θ_{VL}	$CI^{hi}_{\theta_{VL}}$	$CI^{lo}_{\theta_{CV}}$	θ_{CV}	CI^{hi}_{CV}
US	1991	0.162	0.166	0.169	0.329	0.337	0.345	0.177	0.181	0.186	0.158	0.161	0.165	0.396	0.408	0.419	0.574	0.581	0.587
	1997	0.177	0.181	0.185	0.348	0.357	0.366	0.195	0.199	0.204	0.180	0.184	0.189	0.422	0.435	0.447	0.637	0.646	0.654
	2000	0.173	0.177	0.180	0.340	0.348	0.356	0.190	0.194	0.199	0.177	0.181	0.185	0.410	0.421	0.432	0.633	0.643	0.653
	2004	0.179	0.183	0.186	0.361	0.371	0.380	0.197	0.202	0.206	0.178	0.182	0.185	0.439	0.452	0.464	0.625	0.633	0.640
	2007	0.185	0.188	0.191	0.363	0.370	0.377	0.204	0.208	0.212	0.188	0.192	0.196	0.445	0.456	0.466	0.653	0.661	0.669
	2010	0.193	0.197	0.201	0.402	0.411	0.421	0.215	0.219	0.224	0.189	0.192	0.196	0.494	0.508	0.522	0.639	0.646	0.652
DE	1994	0.088	0.095	0.102	0.175	0.188	0.200	0.093	0.100	0.107	0.090	0.097	0.104	0.191	0.207	0.222	0.437	0.456	0.475
	2000	0.088	0.093	0.098	0.174	0.185	0.196	0.092	0.098	0.103	0.090	0.095	0.099	0.190	0.203	0.216	0.439	0.451	0.463
	2004	0.098	0.106	0.114	0.184	0.203	0.222	0.103	0.112	0.121	0.103	0.111	0.119	0.204	0.226	0.248	0.480	0.500	0.519
	2007	0.102	0.111	0.120	0.193	0.210	0.226	0.107	0.117	0.127	0.108	0.118	0.129	0.213	0.234	0.254	0.493	0.522	0.551
	2010	0.103	0.110	0.117	0.198	0.212	0.225	0.109	0.116	0.124	0.107	0.114	0.122	0.221	0.238	0.254	0.482	0.504	0.525

Note. Data from Luxembourg Income Study.