

Karoly, Lynn A.; Schröder, Carsten

**Working Paper**

## Fast methods for jackknifing inequality indices

SOEPpapers on Multidisciplinary Panel Data Research, No. 643

**Provided in Cooperation with:**

German Institute for Economic Research (DIW Berlin)

*Suggested Citation:* Karoly, Lynn A.; Schröder, Carsten (2014) : Fast methods for jackknifing inequality indices, SOEPpapers on Multidisciplinary Panel Data Research, No. 643, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/96415>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

## SOEPpapers

on Multidisciplinary Panel Data Research

SOEP – The German Socio-Economic Panel Study at DIW Berlin

643-2014

# Fast Methods for Jackknifing Inequality Indices

Lynn A. Karoly and Carsten Schröder

## **SOEPpapers on Multidisciplinary Panel Data Research** at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at  
**<http://www.diw.de/soeppapers>**

### **Editors:**

Jürgen **Schupp** (Sociology)

Gert G. **Wagner** (Social Sciences, Vice Dean DIW Graduate Center)

Conchita **D'Ambrosio** (Public Economics)

Denis **Gerstorff** (Psychology, DIW Research Director)

Elke **Holst** (Gender Studies, DIW Research Director)

Frauke **Kreuter** (Survey Methodology, DIW Research Professor)

Martin **Kroh** (Political Science and Survey Methodology)

Frieder R. **Lang** (Psychology, DIW Research Professor)

Henning **Lohmann** (Sociology, DIW Research Professor)

Jörg-Peter **Schräpler** (Survey Methodology, DIW Research Professor)

Thomas **Siedler** (Empirical Economics)

C. Katharina **Spieß** (Empirical Economics and Educational Science)

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)  
DIW Berlin  
Mohrenstrasse 58  
10117 Berlin, Germany

Contact: Uta Rahmann | [soeppapers@diw.de](mailto:soeppapers@diw.de)

# Fast Methods for Jackknifing Inequality Indices

---

Lynn A. Karoly, RAND

Carsten Schröder<sup>^</sup>, SOEP/DIW Berlin and Freie Universität Berlin

**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 develops fast algorithms for jackknifing 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

---

<sup>^</sup> SOEP/DIW Berlin and Freie Universität Berlin. Phone: +49 (0)30 89789-284. Email: [cschroeder@diw.de](mailto:cschroeder@diw.de). We would like to thank Stephen Jenkins for his most helpful comments. Further, we would like to thank Friedrich Bergmann and Jan Krause from University of Kiel 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.<sup>1</sup> 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 central advantage of the jackknife over other resampling methods such as the bootstrap is that it allows the replication of results. A disadvantage of standard jackknife procedures is that for large sample sizes the computational burden is substantial. This is because there are as many subsets as there are observations in the sample, and for each subset the jackknife statistic needs to be computed. This paper offers a solution. We provide fast algorithms, requiring only a few passes through the data, for jackknifing several popular inequality indices: coefficient of variation, variance of the logarithms, mean log deviation, Theil index, and Atkinson index.<sup>2</sup> Since the number of passes is independent of the number of observations, even for large samples the computational burden remains small.

To get an idea of the computational burden see Figure 1. It charts the computer time in minutes for a standard procedure for jackknifing inequality indices as a function of sample size.<sup>3</sup> Computer time increases exponentially in sample size, and for a sample size of about 80,000 cases it already exceeds four hours. Since many comparative inequality analyses rely on data from several points in time, countries and income concepts, computing the jackknife for all results can easily take days or weeks. This is a serious limitation, especially for researchers who use data stored on external servers (e.g., the Luxembourg Income Study) and face limited processing power.

---

<sup>1</sup> For the theoretical justification for the jackknife and other related resampling techniques see Efron (1982).

<sup>2</sup> Algorithms for the Gini coefficient are provided in Karagiannis and Kovacevic (2000) and Yitzhaki (1991). Karoly (1989) 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. 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.

<sup>3</sup> We have used the STATA software package “inequal7.ado” on the following hardware: 64-bit system; 8 GB ram; core(TM)2 Duo CPU; 3GHz.

**Figure 1 about here**

Section 2 explains our jackknife algorithm. 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,  $\mathbf{y} = (y_1, y_2, \dots, y_N)$  and sampling weights,  $\omega_1, \omega_2, \dots, \omega_N$ . Let  $\theta = \theta(\mathbf{y})$  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  $i$ th observation has been deleted.

Following Wolter (1985), the jackknife estimate of the standard error of  $\theta$  is,

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

with  $\bar{\omega} = \frac{1}{N} \sum_{i=1}^N \omega_i$ .<sup>4</sup> 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  $\theta$ ,  $N$ , arithmetic or geometric mean) and a subset-specific correction factor that can be derived with a single run through the data. The

---

<sup>4</sup> An alternative method is to compute the squared differences between the jackknife statistics and their mean (see, for example, Yitzhaki, 1991).

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  $x_i = \ln(y_i)$ .

## 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  $i$ th 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  $\theta_T$ . 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,

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

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

## 2.2 Efficient jackknife procedure for the variance of logarithms

Applying Bessel's correction<sup>5</sup>, 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  $i$ th 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  $i$ th 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):

---

<sup>5</sup> Bessel's correction, the division in the variance formula by  $N-1$  instead of by  $N$ , secures unbiasedness.



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

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

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

Equation (12) can be rewritten as:

$$(13) \quad \begin{aligned} \theta_{VL} &= \underbrace{\frac{1}{N-1} \sum_{j \neq i}^N w_j (x_j - \bar{x}_{(i)})^2}_C + \underbrace{\frac{2}{(N-1)} \sum_{j \neq i}^N w_j ((x_j - \bar{x}_{(i)}) \left( \frac{w_i}{N} \right) (\bar{x}_{(i)} - x_i))}_D \\ &\quad + \underbrace{\frac{1}{N-1} \sum_{j \neq i}^N w_j \left( \frac{w_i}{N} \bar{x}_{(i)} - \frac{w_i}{N} x_i \right)^2}_E + \frac{1}{N-1} w_i (x_i - \bar{x})^2 \end{aligned}$$

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

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

The  $E$  -term after some algebra becomes,

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

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  $\theta_{VL}$ ) and the characteristics of the observation that is left out,  $w_i$  and  $x_i$ . Thus, after having calculated  $N, \bar{x}$ , and  $\theta_{VL}$  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,  $\theta_{A_\varepsilon}$  (with inequality aversion parameter  $\varepsilon = 1$  and  $\varepsilon = 2$ ), the mean log deviation,  $\theta_{MLD}$ , and the coefficient of variation,  $\theta_{CV}$ , are as follow:

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

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

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

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

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.<sup>6</sup> 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,  $CI_{\theta}^{lo}$  and  $CI_{\theta}^{hi}$ , 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.<sup>7</sup>

---

<sup>6</sup> The LIS data for Germany are based on the German Socio-Economic Panel Study (SOEP).

<sup>7</sup> 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 little.

## 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.

## 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 Statistics, 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)

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

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

From (2<sup>MLD</sup>):

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

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

Substituting  $-N[\theta_{MLD} - \ln(\bar{y})] = \sum_{i=1}^N w_i \ln(y_i)$  from (1<sup>MLD</sup>) 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,  $\varepsilon$ .

### Inequality aversion parameter $\varepsilon = 1$

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

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

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 $\varepsilon = 2$

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

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



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:

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

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

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

## 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  $\theta_V$  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  $\bar{y} = \frac{1}{N} [(N-w_i)\bar{y}_{(i)} + y_i w_i]$  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  $\theta_{VL}$  we can rewrite (7<sup>V</sup>) as:

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

Substituting (5<sup>V</sup>), (6<sup>V</sup>), and (8<sup>V</sup>) in (4<sup>V</sup>) 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  $\theta_{VL}$  we can rewrite (9<sup>V</sup>) 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<sup>V</sup>) 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,

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

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<sup>CV</sup>) 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*;
```

```

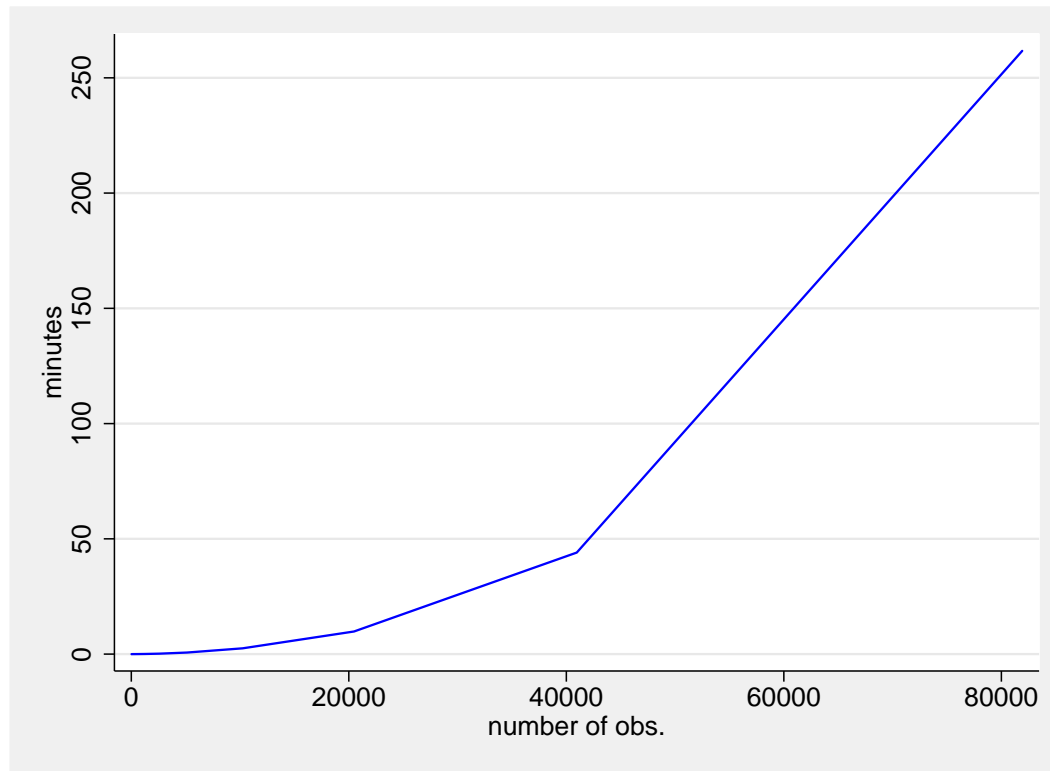
qui gen summand=y/sc_mu*ln(y)*w;
qui sum summand;
qui scalar sc_T=r(mean)-ln(sc_mu);
qui drop summand;
*Variance of log yomes: stored in scalar sc_V*;
qui gen summands=(logy-log(sc_gmu))^2*w;
qui sum summands;
qui scalar sc_VL=r(sum)/(sc_N-1);
qui drop summands;
*Variance and coeff of var: stored in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

Year	Atkinson $\epsilon = 1$			Atkinson $\epsilon = 2$			Mean log deviation			Theil index			Variance of logs		Coeff. of variation			
	$CI_{\theta_{A_1}}^{lo}$	$\theta_{A_1}$	$CI_{\theta_{A_1}}^{hi}$	$CI_{\theta_{A_2}}^{lo}$	$\theta_{A_2}$	$CI_{\theta_{A_2}}^{hi}$	$CI_{\theta_{MLD}}^{lo}$	$\theta_{MLD}$	$CI_{\theta_{MLD}}^{hi}$	$CI_{\theta_T}^{lo}$	$\theta_T$	$CI_{\theta_T}^{hi}$	$CI_{\theta_{VL}}^{lo}$	$\theta_{VL}$	$CI_{\theta_{VL}}^{hi}$	$CI_{\theta_{CV}}^{lo}$	$\theta_{CV}$	$CI_{\theta_{CV}}^{hi}$
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.