Biewen, Martin; Jenkins, Stephen P.

Working Paper

Estimation of Generalized Entropy and Atkinson Inequality Indices from Complex Survey Data

DIW Discussion Papers, No. 345

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Biewen, Martin; Jenkins, Stephen P. (2003) : Estimation of Generalized Entropy and Atkinson Inequality Indices from Complex Survey Data, DIW Discussion Papers, No. 345, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
http://hdl.handle.net/10419/18081

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.
Martin Biewen and Stephen P. Jenkins

Estimation of Generalized Entropy and Atkinson Inequality Indices from Complex Survey Data

Berlin, May 2003
Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.
Estimation of Generalized Entropy and Atkinson Inequality Indices from Complex Survey Data

Martin Biewen
University of Mannheim
IZA, Bonn
DIW Berlin

Stephen P. Jenkins
ISER, University of Essex
IZA, Bonn
DIW Berlin

This version: 11 April 2003

Abstract. Applying a method suggested by Woodruff (1971), we derive the sampling variances of Generalized Entropy and Atkinson inequality indices when estimated from complex survey data. It turns out that this method also greatly simplifies the calculations for the i.i.d. case when compared to previous derivations in the literature. Both cases are illustrated with examples from the German Socio-Economic Panel Study and the British Household Panel Survey.

JEL-Classification: C14, D31
Keywords: Inequality, Statistical Inference, Complex Surveys

Correspondence:
Martin Biewen, Department of Economics, University of Mannheim, Verfügungsgebäude L7, 3-5, 68131 Mannheim, Germany, Fax: +49-621-1811841, biewen@rhein.vwl.uni-mannheim.de

1Financial support from the the Deutsche Forschungsgesellschaft (DFG) and core funding to ISER from the ESRC and the University of Essex is gratefully acknowledged.
1 Introduction

Probability weighting, clustering, and stratification, are survey design features underlying much of the survey data that economists and others use. It is well known that these features have a potentially large impact on the sampling variability of statistics computed from such surveys. Nevertheless, they are rarely taken into account in practical work, the measurement of inequality being no exception. We derive estimates for the sampling variance of two commonly-used classes of inequality indices, the Generalized Entropy and the Atkinson family of indices, using the approach of Woodruff (1971). It turns out that Woodruff’s method also greatly simplifies the computation of variance estimates in an i.i.d. framework when compared to previous derivations in the literature. In order to assess the error made by not taking into account clustering and stratification we apply our results to a sample extracted from the German Socio-Economic Panel (GSOEP) and the British Household Panel Survey (BHPS).

2 Estimation from complex surveys

Generalized Entropy and Atkinson inequality measures can be written, as we show below, as functions $I = f(T)$ of population totals $T = (T_1, \ldots, T_K)$. Population total $T_k$, $k = 1, \ldots, K$ is given by the summation of an observational variable $t_{hijk}$ over the different stages of the sampling design, i.e. $T_k = \sum_{n=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{M_i} t_{hijk}$ where $L$ denotes the number of strata, $N_h$ the number of clusters in stratum $h$ and $M_i$ the number of individuals in cluster $i$. If the sampling design involves more than one stage of clustering it suffices to consider the first stage only (see e.g. Cochran, 1977). Replacing totals $T$ by their estimates $\hat{T}$, the index is then estimated as $\hat{I} = f(\hat{T})$ with $\hat{T}_k = \sum_{n=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{m_i} w_{hij} t_{hijk}$ where $n_h$ is the number of actually sampled first stage clusters and $m_i$ the number of actually sampled individuals in cluster $i$. The sampling weight of individual $hij$ is given by $w_{hij}$. Assuming that the sample is large enough that a Taylor approximation of $f(\cdot)$ holds, the variance of $\hat{I}$ can be approximated by the variance of the first order residual $\sum_{k=1}^K (\partial f(T)/\partial T_k) \hat{T}_k$. Woodruff (1971) observed that this variance can be easily determined by reversing the order of summation in the residual, i.e. $\text{var}(\hat{I}) \approx \text{var}(\sum_{n=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_i} w_{hij} \left[ \sum_{k=1}^K (\partial f(T)/\partial T_k) t_{hijk} \right]) = \text{var}(\hat{S})$. Note that $\hat{S}$ is of the

---

2 An alternative but conceptually less straightforward approach to variance estimation in complex surveys is the estimating equations approach described in Binder (1983) and Binder and Patak (1994). It turns out that this approach leads to the same estimators derived here. Calculations are available from the authors on request.
same form as the $T_k$'s so that the problem is reduced to the estimation of the sampling variance of a total estimator for which well-known formulas exist (see Cochran, 1977, or Deaton, 1997). Using these formulas (and replacing $T$ by $\hat{T}$ in the derivative), the variance estimate for $\hat{I}$ is

$$\text{var}(\hat{I}) = \frac{L}{n_h} - \frac{\hat{I}}{n_h - 1} \left( \sum_{i=1}^{m_i} w_{hij} \hat{s}_{hij} - \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_i} w_{hij} \hat{s}_{hij}}{n_h} \right)^2$$

(1)

with $\hat{s}_{hij} = \sum_{k=1}^{K} (\partial f(\hat{T})/\partial \hat{T}) t_{hijk}$.

If $y_{hij}$ represents the income of individual $hij$, then population Generalized Entropy and Atkinson indices are given by

$$I_{GE}^P = (\alpha^2 - \alpha)^{-1} \left[ U_0^{-\alpha} U_1^{1-\alpha} \alpha - 1 \right], \quad \alpha \in \mathbb{R} \setminus \{0, 1\}$$

(2)

$$I_{theil} = T_{1,1} U_1^{-1} - \log(U_1 U_0^{-1}), \quad \alpha \to 1$$

(3)

$$I_{MLD} = -T_{0,1} U_0^{-1} \log(U_1 U_0^{-1}), \quad \alpha \to 0$$

(4)

$$I_{\varepsilon}^A = 1 - \frac{U_0^{-\varepsilon/(1-\varepsilon)} U_1^{1/(1-\varepsilon)}}{U_1^{1-\varepsilon}}, \quad \varepsilon \geq 0, \quad \varepsilon \neq 1,$$

(5)

$$I_{\varepsilon}^A = 1 - U_0 U_1^{-1} \exp(T_{0,1} U_0^{-1}), \quad \varepsilon \to 1$$

(6)

for totals $U_\alpha = \sum_{n=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_i} (y_{hij})^\alpha$ and $T_\alpha = \sum_{n=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_i} (y_{hij})^\alpha (\log y_{hij})$. Note that $U_0$ is the population size. Estimates of these indices are obtained by replacing $U_\alpha$, $T_\alpha$ with $\hat{U}_\alpha = \sum_{n=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{m_i} w_{hij} (y_{hij})^\alpha$ and $\hat{T}_\alpha = \sum_{n=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{m_i} w_{hij} (y_{hij})^\alpha (\log y_{hij})$.

The corresponding variances are then given by substituting $\hat{s}_{hij}$ in (1) with

$$\hat{s}_{hij}^{GE} = \frac{1}{\alpha} \hat{U}_0 \hat{U}_1 \hat{U}_0^{-\alpha} \hat{U}_1^{1-\alpha} \hat{U}_0^{-1} \cdot y_{hij} + \frac{1}{\alpha^2 - \alpha} \hat{U}_0^{-\alpha} \hat{U}_1^{1-\alpha} \cdot (y_{hij})^\alpha$$

(7)

$$\hat{s}_{hij}^{Theil} = \hat{U}_1^{1-\alpha} \cdot (y_{hij} \log y_{hij} - \hat{U}_1^{1} \hat{U}_1^{1-\alpha} (\hat{T}_{1,1} \hat{U}_1^{1-\alpha} + 1) \cdot y_{hij} + \hat{U}_1^{1-\alpha})$$

(8)

$$\hat{s}_{hij}^{MLD} = -\hat{U}_0^{-1} \cdot \log y_{hij} + \hat{U}_1 \cdot y_{hij} + \hat{U}_0^{-1} \left( \hat{T}_{0,1} \hat{U}_0^{1-\alpha} - 1 \right)$$

(9)

$$\hat{s}_{hij}^{A} = \frac{\varepsilon}{1-\varepsilon} \hat{U}_1^{1-\alpha} \hat{U}_0^{-\alpha} \hat{U}_0^{1-\varepsilon} \cdot y_{hij} + \frac{1}{1-\varepsilon} \hat{U}_0^{-\alpha} \hat{U}_1^{1-\alpha} \cdot (y_{hij})^{1-\varepsilon}$$

(10)

$$\hat{s}_{hij}^{A} = \left( \hat{I}_A - 1 \right) \hat{U}_0 \cdot (1 - \hat{U}_0 \hat{T}_{0,1}) + \left( 1 - \hat{I}_A \right) \hat{U}_1 \cdot y_{hij} + \left( \hat{I}_A - 1 \right) \hat{U}_0^{-1} \cdot \log y_{hij}$$

(11)

These variance estimators allow arbitrary correlations between observations below the first-stage clusters. They are therefore an alternative to the estimators developed by Schluter and Trede (2002).
3 Application to the i.i.d. case

In an i.i.d. framework the above indices are usually treated as follows. Income $x_i$ and weight $w_i$ of observation $i = 1, \ldots, n$ are regarded as i.i.d. draws from a population $(x, w)$.

The index in question can then be represented as a function $I = g(\mu)$ of population moments $\mu = E(X_i)$, where $X_i$ is a vector-valued function of $(x_i, w_i)$. It is estimated as $\hat{I} = g(\hat{X})$ and its sampling variance as $n^{-1} \nabla g(\hat{X}) / \nabla \hat{X} V g(\hat{X})$ (Cowell, 1989, or for the case without weights, Thistle, 1990). By contrast, Woodruff’s method would yield a variance estimate $n^2 \var(X_i)$.

It is easy to see that both estimates are identical. However, Woodruff’s method leads to much simpler expressions as the problem is reduced to estimating the sampling variance of a scalar. In particular, no covariances need to be computed. Defining $\hat{\mu}_0 = E(w_i x_i^\alpha)$, $\tau_0 = E(w_i x_i^\alpha \log x_i)$, $\hat{\tau}_{\alpha, \gamma} = n^{-1} \sum_{i=1}^n w_i x_i^\alpha (\log x_i)$, the sampling variances of the indices

\begin{equation}
I_{GE}^\alpha = (\alpha^2 - \alpha)^{-1} \left[ \mu_0^{\alpha - 1} \mu_1^{-\alpha} \mu_\alpha - 1 \right], \quad \alpha \in \mathbb{R} \setminus \{0, 1\}
\end{equation}

\begin{equation}
I_{Theil}^\alpha = \tau_{1,1} \mu_1^{-1} \log(\mu_1 \mu_0^{-1}), \quad \alpha \to 1
\end{equation}

\begin{equation}
I_{MLD}^\alpha = -\tau_{0,1} \mu_0^{-1} \log(\mu_1 \mu_0^{-1}), \quad \alpha \to 0
\end{equation}

\begin{equation}
I_A^\epsilon = 1 - \mu_0^{-\epsilon/(1-\epsilon)} \mu_1^{-1} \mu_1^{1/(1-\epsilon)}, \quad \epsilon \geq 0,
\end{equation}

\begin{equation}
I_A^1 = 1 - \mu_0 \mu_1^{-1} \exp(\tau_{0,1} \mu_0^{-1}), \quad \epsilon \to 1
\end{equation}

can therefore simply be estimated by substituting

\begin{equation}
\hat{z}_{i,GE} = \frac{1}{\alpha} \hat{\mu}_0 \hat{\mu}_1 \alpha \hat{\mu}_0^{-2} - \frac{1}{\alpha - 1} \hat{\mu}_0 \hat{\mu}_1^{-a-1} \hat{\mu}_0^{a-1} x_i + \frac{1}{\alpha^2 - \alpha} \hat{\mu}_0^{a-1} \hat{\mu}_1^{-\alpha} x_i^\alpha_x
\end{equation}

\begin{equation}
\hat{z}_{i,Theil} = \hat{\mu}_1^{-1} \log x_i - \hat{\mu}_1^{-1} (\hat{\tau}_{1,1} \hat{\mu}_1^{-1} + 1) \cdot x_i + \hat{\mu}_0^{-1}
\end{equation}

\begin{equation}
\hat{z}_{i,MLD} = -\hat{\mu}_0^{-1} \log x_i + \hat{\mu}_1^{-1} \cdot x_i + \mu_0^{-1} (\hat{\tau}_{0,1} \hat{\mu}_0^{-1} - 1)
\end{equation}

\begin{equation}
\hat{z}_i^A = \frac{\epsilon}{1 - \epsilon} \hat{\mu}_1 \hat{\mu}_1^{-1} \mu_0^{-1} \mu_1^{-2} + \mu_0^{-1} \hat{\mu}_1^{-1} \mu_1^{-2} \cdot x_i - \frac{1}{1 - \epsilon} \mu_0^{-1} \hat{\mu}_1^{-1} \mu_1^{-2} \cdot x_i^{1-\epsilon}
\end{equation}

\begin{equation}
\hat{z}_i^1 = (\hat{I}_A - 1) \hat{\mu}_0^{-1} \left( 1 - \hat{\mu}_0^{-1} \hat{\tau}_{0,1} \right) + (1 - \hat{I}_A) \hat{\mu}_0^{-1} \cdot x_i + (\hat{I}_A - 1) \hat{\mu}_0^{-1} \cdot \log x_i
\end{equation}

for $\hat{z}_i$ in

\begin{equation}
\var(\hat{I}) = \frac{1}{n(n - 1)} \sum_{i=1}^n w_i \hat{z}_i - \frac{\sum_{i=1}^n w_i \hat{z}_i}{n}.
\end{equation}

Equation (22) has a similar structure to equation (1), but note that the weights are treated differently in the complex survey and i.i.d. cases.\footnote{If the distribution of household income among individuals is analyzed, then observational units $i$ are households and observations are replicated at the household level. In this case weights are $w_i = w'_i w''_i$ with household size $w'_i$ and sample weight $w''_i$. See Biewen (2002).} 

\footnote{For more discussion of different weighting concepts, see Cowell and Jenkins (2003).}
4 Empirical illustration

We contrast the variance estimators using data from the first waves of the German Socio-
Economic Panel (SOEP Group, 2001), and the British Household Panel Survey (Taylor et al.,
2002). The GSOEP and the BHPS are widely used to analyze the income distribution in these
two countries. Moreover, they provide information on primary sampling units (clusters) and
strata identification variables. The first wave of each survey was chosen to avoid complications
of the panel design. We considered the distribution of income among individuals. Following con-
vention, each person was assumed to receive the equivalent household income of the household
to which she belonged. (The equivalence scale was the square root of household size.)

The survey estimates shown in column (1) of Table 1 take into account sampling weights,
stratification and clustering. Replication of observations at the household level is automatically
accounted for, as this represents a form of clustering below the first stage clusters. The i.i.d.
estimates shown in column (2) only take into account sampling weights and replication of ob-
servations at the household level, but not stratification or clustering. A comparison of columns
(1) and (2) indicates that ignoring clustering and stratification makes surprisingly little dif-
ference for these data sets. By contrast, the estimates shown in columns (3) and (4) suggest
that taking into account the replication of observations at the household level is much more
important. The results in column (3) ignore the replication of observations at the household
level, whereas those in column (4) take it into account. (Both (3) and (4) ignore first-stage clu-
stering and stratification.) For the German data, this leads to standard errors that are about
twice as large. This shows that survey design can matter. However, the precise effect appears
to depend on the survey analysed: corresponding estimates in columns (3) and (4) differ little
when BHPS data are used, by contrast with the GSOEP case.

— Table 1 near here —

The fact that ignoring stratification and first-stage clustering has only a small impact
might be interpreted as good news for practitioners using these data, or for those using surveys
in which primary sampling unit and strata identification variables are not made available, but
it is not clear that this empirical finding can be generalized. Whatever the case, our variance
estimators provide a straightforward means by which researchers can accommodate a range of
design effects in their analysis.

5 Stata programs to compute the estimators are available from the authors on request.
5 References


## 6 Tables

**Table 1. Income inequality\(^1\) in West Germany (1984) and Britain (1991)**

<table>
<thead>
<tr>
<th>Index</th>
<th>(1) Survey(^2)</th>
<th>(2) I.i.d.(^3)</th>
<th>(3) Survey, i.i.d.(^4)</th>
<th>(4) Survey(^5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>std. err.</td>
<td>estimate</td>
<td>std. err.</td>
</tr>
<tr>
<td><strong>German Socio-Economic Panel</strong> (1984)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE(-1)</td>
<td>0.4397</td>
<td>0.2971</td>
<td>0.3992</td>
<td>0.2573</td>
</tr>
<tr>
<td>MLD</td>
<td>0.1339</td>
<td>0.0153</td>
<td>0.1338</td>
<td>0.0136</td>
</tr>
<tr>
<td>Theil</td>
<td>0.1540</td>
<td>0.0276</td>
<td>0.1540</td>
<td>0.0246</td>
</tr>
<tr>
<td>GE(2)</td>
<td>0.3673</td>
<td>0.1586</td>
<td>0.3502</td>
<td>0.1400</td>
</tr>
<tr>
<td>Atkinson(0.5)</td>
<td>0.0658</td>
<td>0.0079</td>
<td>0.0661</td>
<td>0.0071</td>
</tr>
<tr>
<td>Atkinson(1)</td>
<td>0.1253</td>
<td>0.0134</td>
<td>0.1253</td>
<td>0.0119</td>
</tr>
<tr>
<td>Atkinson(1.5)</td>
<td>0.2153</td>
<td>0.0477</td>
<td>0.2101</td>
<td>0.0419</td>
</tr>
<tr>
<td>Atkinson(2)</td>
<td>0.4679</td>
<td>0.1682</td>
<td>0.4439</td>
<td>0.1591</td>
</tr>
<tr>
<td><strong>British Household Panel Survey</strong> (1991)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE(-1)</td>
<td>0.3656</td>
<td>0.0605</td>
<td>0.3627</td>
<td>0.0541</td>
</tr>
<tr>
<td>MLD</td>
<td>0.1907</td>
<td>0.0051</td>
<td>0.1914</td>
<td>0.0050</td>
</tr>
<tr>
<td>Theil</td>
<td>0.1779</td>
<td>0.0050</td>
<td>0.1784</td>
<td>0.0050</td>
</tr>
<tr>
<td>GE(2)</td>
<td>0.2064</td>
<td>0.0085</td>
<td>0.2071</td>
<td>0.0086</td>
</tr>
<tr>
<td>Atkinson(0.5)</td>
<td>0.0873</td>
<td>0.0022</td>
<td>0.0875</td>
<td>0.0022</td>
</tr>
<tr>
<td>Atkinson(1)</td>
<td>0.1736</td>
<td>0.0042</td>
<td>0.1741</td>
<td>0.0041</td>
</tr>
<tr>
<td>Atkinson(1.5)</td>
<td>0.2685</td>
<td>0.0084</td>
<td>0.2693</td>
<td>0.0082</td>
</tr>
<tr>
<td>Atkinson(2)</td>
<td>0.4223</td>
<td>0.0404</td>
<td>0.4204</td>
<td>0.0363</td>
</tr>
</tbody>
</table>

---

\(^1\) Income refers to monthly equivalent household net income distributed among individuals (equivalence scale = square root of household size).

\(^2\) Survey estimator, individual data, weight = individual sample weight, accounting for clustering and stratification; replication of observations at household level automatically accounted for.

\(^3\) I.i.d. estimator, household data, weight = household size \(\times\) household sample weight (thus accounting for replication of observations at household level), ignoring clustering and stratification.

\(^4\) Survey estimator, individual data, weight = individual sample weight, ignoring clustering, stratification and replication of observations at household level (identical to i.i.d. estimator, individual data, weight = individual sample weight).

\(^5\) Survey estimator, individual data, weight = individual sample weight, ignoring clustering and stratification but accounting for replication of observations at household level (households are interpreted as clusters).

---

\(^6\) 110 strata, 516 clusters, 4232 households, 9441 individuals.

\(^7\) 75 strata, 250 clusters, 4814 households, 11616 individuals.