Variance estimation for Generalized Entropy and Atkinson indices: the complex survey data case

Martin Biewen
University of Frankfurt/Main
and
Stephen P. Jenkins
Institute for Social & Economic Research
University of Essex
Email: stephenj@essex.ac.uk

Presentation at UK Stata User Group meeting, London, 17–18 May 2005

Inequality indices: specialist measures of the dispersion of a distribution

Imposition of a small number of axioms, substantially restricts the functional form that indices may have.

Axioms for $I(\boldsymbol{y})$:

- Anonymity (a.k.a. symmetry): $I(\boldsymbol{y})$ depends on $\boldsymbol{y}$ only
- Principle of Transfers: a mean-preserving spread in $\boldsymbol{y}$ increases $I(\boldsymbol{y})$
- Scale invariance: $I(k\boldsymbol{y}) = I(\boldsymbol{y})$ for all scalar $k > 0$
- Replication invariance: $I(\boldsymbol{y}, \boldsymbol{y}, \ldots, \boldsymbol{y}) = I(\boldsymbol{y})$
- Normalization: $I(\boldsymbol{y}) = 0$ if $\boldsymbol{y} = \mu$

Classes of inequality measures satisfying the axioms

Generalized Entropy (transfer sensitivity parameter α)

$$I_{\mathrm{GE}}^{\alpha}(F) := \frac{1}{\alpha^2-\alpha}\int \left[\left[\frac{x}{\mu(F)}\right]^{\alpha} - 1\right] dF(x)$$

$\alpha \neq 0, 1$

$= CV^2/2$ if $\alpha = 2$

$$I_{\mathrm{Theil}}(F) := \int \frac{x}{\mu(F)} \log\left(\frac{x}{\mu(F)}\right) dF(x)$$

$\alpha \to 1$

$$I_{\mathrm{MLD}}(F) := -\int \log\left(\frac{x}{\mu(F)}\right) dF(x)$$

$\alpha \to 0$

Atkinson (inequality aversion parameter ε > 0

For each member of I_A, there is an ordinally equivalent member of I_{GE}

$$I_{\mathrm{A}}^{\varepsilon}(F) := 1 - \frac{1}{\mu(F)}\left[\int x^{1-\varepsilon} dF(x)\right]^{\frac{1}{1-\varepsilon}}$$

Gini coefficient

$$I_{\mathrm{Gini}}(F) \quad : \quad = \frac{1}{2\mu(F)} \int\int |x - x'| \, dF(x) dF(x') \quad = \quad 1 - 2\int_0^1 L(F;q)dq$$

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

Formulae from Cowell (2000)

Estimation of inequality indices

- These indices are routinely calculated by many analysts …
  - The most commonly-used programs among Stata users are **`ineqdeco`** and **`inequal7`** (available using **`ssc`**)
- **But** only rarely do analysts report estimates of the associated sampling variances (SEs) of the estimates
  - Analytical derivations to date have omitted some important situations (and indices)
    - Most assume i.i.d. observations (cf. survey clustering or other sample dependencies!), and don't consider probability weighting (cf. stratification!)
  - The methods that do exist are not 'well known'
  - Lack of available software
    - But cf. **`geivars`** (Cowell 1988, linearization methods; i.i.d. assumptions) and **`ineqerr`** (bootstrap), both available using **`ssc`**

What we provide

- Estimates of indices and associated sampling variances for all members of the GE and Atkinson classes, while also …
- Accounting for clustering and stratification, and for the i.i.d. case
- Analytical results (see our paper) and new Stata programs (version 8.2): **`svygei`** and **`svyatk`**
- Based on Taylor-series linearization methods combined with a result from Woodruff (*JASA*, 1971)
  - Standard linearization methods stymied because indices are (functions of) moments in addition to means (cf. poverty)
  - Results don't apply to Gini index or other measures based on order statistics

Overview of analytical derivation

- Write the estimator of each index as a function of population totals (involves sums over clusters, weights, etc.)
- Assuming *N* sufficiently large that 1st order Taylor series approximation holds, then the variance of each estimator is well approximated by the variance of the first order 'residual' for the index
- As is, each expression is not easily calculated, but …
- (Woodruff): reversing the order of summation in the 'residual' expression $\Rightarrow$ estimation is equivalent to derivation of a sampling variance of a total estimator for which one can apply standard **svy** methods

The programs: `svygei`, `svyatk`

`svygei varname [if exp] [in range] [, alpha(#) subpop(varname) level(#)`

Calculations for $\alpha = -1, 0, 1, 2, 3$ (use **`alpha(#)`** option to choose one α other than 3)

`svyatk varname [if exp] [in range] [, epsilon(#) subpop(varname) level(#)`

Calculations for $\varepsilon = 0.5, 1, 1.5, 2, 2.5$ (use **`epsilon(#)`** option to choose one ε other than 2.5)

where, of course, the data have first been **`svyset.`**

- How the data are organised, and described using **`svyset`**, is of crucial importance …

Selected examples of survey data set-up for estimation of inequality among individuals

1. Observation unit is person; sampling unit is household; all persons in each household attributed with the income of the household to which they belong; individual sample weight available (‘**xewght**’), but no information about PSU or strata

   ```
   svyset [pw = xewght], psu(hh_id)
   ```

2. As (1), except also know PSU and strata information (includes allowance for within-household correlation):

   ```
   svyset [pw = xewght], psu(PSUid) strata(STRATAid)
   ```

3. Observation unit is household; sampling unit is household; weight = household sample weight × household size (‘**xhhwt**’), but no information about PSU or strata

   ```
   svyset [pw = xhhwt]
   ```

   i.i.d. case

Illustration

- British Household Panel Survey, wave 11 data (2001) used as a cross-section
- 9,979 individuals in 4,058 households (‘**`hid`**’); 250 PSUs (‘**`psu`**’), 75 strata (‘**`strata`**’).
- Needs-adjusted post-tax post-benefit household income (‘**`net`**’)
- Each individual attributed with the income of his/her household ($\Rightarrow$ ‘clustering’ within households)
  - Even if survey does not include PSU and strata identifiers, you should account for this (use household identifier as PSU variable)

Generalized Entropy indices

```
. svyset [pweight =  xewght], psu(psu) strata(strata)

. svygei net

Complex survey estimates of Generalized Entropy inequality indices

pweight: xewght                               Number of obs     = 9779
Strata: strata                                Number of strata = 75
PSU: psu                                      Number of PSUs    = 250
                                              Population size   = 9765.8343
---------------------------------------------------------------------------
Index     |  Estimate   Std. Err.      z      P>|z|      [95% Conf. Interval]
---------+-----------------------------------------------------------------
GE(-1)    |  .3132977   .03751986     8.35    0.000      .2397601   .3868353
MLD       |  .1742045   .00608278    28.64    0.000      .1622825   .1861266
Theil     |  .1676984   .00755704    22.19    0.000      .1528869   .1825099
GE(2)     |   .211649   .01868139    11.33    0.000      .1750341   .2482638
GE(3)     |  .3841949   .07587589     5.06    0.000      .2354809    .532909
---------------------------------------------------------------------------

. ineqerr net [w = xewght], reps(100) psu(psu)
<snip>
Variable |    Reps    Observed       Bias    Std. Err.    [95% Conf. Interval]
---------+-------------------------------------------------------------------
   Theil |     100    .1676984   .0010148    .0113708    .1451364   .1902605   (N)
<snip>
```

Bootstrap (100 reps): larger SE. Estimation time = 25.7 secs (cf. 0.89 secs)

Atkinson indices

```
. svyset [pweight =  xewght], psu(psu) strata(strata)

. svyatk net

Complex survey estimates of Atkinson inequality indices

pweight: xewght                                Number of obs    = 9779
Strata: strata                                 Number of strata = 75
PSU: psu                                       Number of PSUs   = 250
                                               Population size  = 9765.8343
---------------------------------------------------------------------------
Index     |  Estimate   Std. Err.       z      P>|z|     [95% Conf. Interval]
---------+-----------------------------------------------------------------
A(0.5)    |  .0808326   .00291639    27.72    0.000      .0751166   .0865487
A(1)      |   .159875   .00511029    31.28    0.000       .149859    .169891
A(1.5)    |  .2484654   .00896696    27.71    0.000      .2308905   .2660403
A(2)      |   .385219   .02836169    13.58    0.000      .3296311   .4408068
A(2.5)    |   .641532   .07499909     8.55    0.000      .4945365   .7885276
---------------------------------------------------------------------------
```

Sub-population option

```
. ge male = hgsex == 2

. svygei net, subpop(male)

Complex survey estimates of Generalized Entropy inequality indices

pweight: xewght                               Number of obs     = 9779
Strata: strata                                Number of strata = 75
PSU: psu                                      Number of PSUs    = 250
                                              Population size   = 9765.8343
Subpop: male, subpop. size = 5192.4171
---------------------------------------------------------------------------
Index     |  Estimate   Std. Err.      z       P>|z|     [95% Conf. Interval]
---------+-----------------------------------------------------------------
GE(-1)    |  .3031452   .02980789    10.17     0.000      .2447228   .3615676
MLD       |  .1793633   .00789997    22.70     0.000      .1638797    .194847
Theil     |  .1738743   .01083914    16.04     0.000        .15263   .1951186
GE(2)     |  .2252216   .03066442     7.34     0.000      .1651204   .2853227
GE(3)     |  .4414405   .1419052     3.11     0.002      .1633114   .7195695
---------------------------------------------------------------------------
```

Empirical illustration in our paper

- BHPS income data for 2001 (almost identical to above), and
- German Socio-Economic Panel data for 2001 (12,939 persons in 5,195 households; 1,004 PSUs, 169 strata)
  - Inequality larger in Britain than Germany, for all indices, and difference is statistically significant (conventional levels)
  - *z*-ratios (index ÷ SE) vary from 7.5 to 23.9 (DE) and 5.1 to 31.9 (GB), being smallest for very top-sensitive indices and largest for middle-sensitive indices
  - Although sample is larger in Germany, *z*-ratios are not always smaller (reflecting different sample designs)

Empirical illustration (ctd.)

Effects of different assumptions about survey design on sampling variance estimates?

- For each index, the estimated standard error is larger if one accounts for survey clustering and stratification (unsurprising), but …
- Results suggest that accounting for survey design features *per se* have little (additional) effect on variance estimates **as long as** the replication of incomes within multi-person households is accounted for

Conclusions

- Researchers now have the means to estimate sampling variances for most of the inequality indices in common use, accommodating a range of potential assumptions about design effects

Topics for future research:

- GE indices are additively decomposable by population subgroup (**`ineqdeco`**): extend results here to the components of decompositions (cf. **`subpop`** option giving a single within-group estimate)
- Extend results to Gini coefficient and other measures based on order statistics (Lorenz curves etc.)

Selected references

Biewen, M. and Jenkins, S.P. (2003), 'Estimation of Generalized Entropy and Atkinson indices from complex survey data', Working Paper 2003-11, Institute for Social and Economic Research, University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2003-11.pdf, *Oxford Bulletin of Economics and Statistics*, submitted.

Cowell, F.A. (2000), 'Measurement of inequality', in: A.B. Atkinson and F. Bourguignon (eds), *Handbook of Income Distribution*, *Volume 1*, Elsevier Science, Amsterdam.

Woodruff, R.S. (1971), 'A simple method for approximating the variance of a complicated estimate', *Journal of the American Statistical Association*, 66, 411–4.