Variance estimation by linearization
for indicators of poverty and social exclusion
in a person and household survey context
Illustration with swiss SILC 2009

Eric Graf

Statistical Institute
University of Neuchâtel, Switzerland
www.unine.ch/statistics

NTTS, 7th of March 2013
Content

Introduction
Content

Introduction

Indicators for poverty and social exclusion
  - Empirical calculations
  - Parametric calculations
  - GB2 distribution and inequality measures

Linearization techniques

Illustrative results

Conclusions
Content

Introduction

Indicators for poverty and social exclusion
  Empirical calculations
  Parametric calculations
  GB2 distribution and inequality measures

Linearization techniques
  Generalized linearization
  Influence functions for poverty and inequality measure
    Estimation of income density function
    Results from simulations
  Variance estimation of indicators under GB2
  Calibrated survey weights and linearization
Content

Introduction

Indicators for poverty and social exclusion
  Empirical calculations
  Parametric calculations
  GB2 distribution and inequality measures

Linearization techniques
  Generalized linearization
  Influence functions for poverty and inequality measure
    Estimation of income density function
    Results from simulations
  Variance estimation of indicators under GB2
  Calibrated survey weights and linearization

Illustrative results
Content

Introduction

Indicators for poverty and social exclusion
  Empirical calculations
  Parametric calculations
  GB2 distribution and inequality measures

Linearization techniques
  Generalized linearization
  Influence functions for poverty and inequality measure
    Estimation of income density function
    Results from simulations
  Variance estimation of indicators under GB2
  Calibrated survey weights and linearization

Illustrative results

Conclusions
Introduction : variance estimation

European projects

- **DACSEIS** (2001-2004, FP5) *Data Quality in Complex Surveys within the New European Information Society*
- **KEI** (2004-2008, FP6) *Knowledge Economy Indicators*
- **SAMPLE** (2008-2011, FP7) *Small Area Methods for Poverty and Living Condition Estimates* ↔ EU-SILC
- **AMELI** (2008-2011, FP7) *Advanced Methodology for European Laeken Indicators* : estimation, variance estimation, robustness, ... ↔ EU-SILC
- **Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys** : recommendations, precision requirements, best practices on variance estimation, computing standard errors for national and european statistics, references, ...
Introduction: variance estimation

European projects

- **DACSEIS** (2001-2004, FP5) *Data Quality in Complex Surveys within the New European Information Society*
- **KEI** (2004-2008, FP6) *Knowledge Economy Indicators*
- **SAMPLE** (2008-2011, FP7) *Small Area Methods for Poverty and Living Condition Estimates ↔ EU-SILC*
- **AMELI** (2008-2011, FP7) *Advanced Methodology for European Laeken Indicators: estimation, variance estimation, robustness,... ↔ EU-SILC*
- **Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys:** recommendations, precision requirements, best practices on variance estimation, computing standard errors for national and european statistics, references, ...
Introduction: variance estimation

European projects

- **DACSEIS** (2001-2004, FP5) *Data Quality in Complex Surveys within the New European Information Society*
- **KEI** (2004-2008, FP6) *Knowledge Economy Indicators*
- **SAMPLE** (2008-2011, FP7) *Small Area Methods for Poverty and Living Condition Estimates* ↔ EU-SILC
- **AMELI** (2008-2011, FP7) *Advanced Methodology for European Laeken Indicators* : estimation, variance estimation, robustness,... ↔ EU-SILC
- *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys* : recommendations, precision requirements, best practices on variance estimation, computing standard errors for national and european statistics, references, ...
Introduction: variance estimation

European projects

- **DACSEIS** (2001-2004, FP5) *Data Quality in Complex Surveys within the New European Information Society*
- **KEI** (2004-2008, FP6) *Knowledge Economy Indicators*
- **SAMPLE** (2008-2011, FP7) *Small Area Methods for Poverty and Living Condition Estimates* ↔ EU-SILC
- **AMELI** (2008-2011, FP7) *Advanced Methodology for European Laeken Indicators*: estimation, variance estimation, robustness,... ↔ EU-SILC
- **Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys**: recommendations, precision requirements, best practices on variance estimation, computing standard errors for national and european statistics, references, ...
Introduction: variance estimation

European projects

▸ DACSEIS (2001-2004, FP5) *Data Quality in Complex Surveys within the New European Information Society*

▸ KEI (2004-2008, FP6) *Knowledge Economy Indicators*

▸ SAMPLE (2008-2011, FP7) *Small Area Methods for Poverty and Living Condition Estimates* ↔ EU-SILC

▸ AMELI (2008-2011, FP7) *Advanced Methodology for European Laeken Indicators*: estimation, variance estimation, robustness,... ↔ EU-SILC

▸ *Handbook on Precision Requirements and Variance Estimation for ESS Household Surveys*: recommendations, precision requirements, best practices on variance estimation, computing standard errors for national and european statistics, references, ...
Goal

Through a household survey (example: SILC), we collect, among other measures, some income variable, call it $y$.

We intend to do inference to the population in publishing estimated values with confidence intervals for several indicators of poverty and social exclusion.

Press release

in this country, the Gini index is of 29% ± 1.5%.

Need of procedure to estimate the variance.
Sources of variability of a survey based estimator

- Sampling variance: the survey providing the salaries \( y \) originates from a probabilistic sample \( \rightarrow \) need to describe the sample-to-sample variations
- Non-response variance, imputation variance, over-coverage variance, response-variance, ...
- Mostly the weights provided with the collected sample reflect some of the variability sources: sampling design, correction for unit non response, there may have been weight sharing, calibration, etc. The weights are random too!
Sources of variability of a survey based estimator

- Sampling variance: the survey providing the salaries $y$ originates from a probabilistic sample ⇒ need to describe the sample-to-sample variations

- Non-response variance, imputation variance, over-coverage variance, response-variance, ...

- Mostly the weights provided with the collected sample reflect some of the variability sources: sampling design, correction for unit non response, there may have been weight sharing, calibration, etc. The weights are random too!
Sources of variability of a survey based estimator

- Sampling variance: the survey providing the salaries \( y \) originates from a probabilistic sample \( \Rightarrow \) need to describe the sample-to-sample variations
- Non-response variance, imputation variance, over-coverage variance, response-variance, ...
- Mostly the weights provided with the collected sample reflect some of the variability sources: sampling design, correction for unit non response, there may have been weight sharing, calibration, etc. The weights are random too!
Introduction
Indicators for poverty and social exclusion
Linearization techniques
Illustrative results
Conclusions

Additional problems for Laeken indicators (ex. GINI)

- The distribution of income is hard to model, far away from a (log-)Normal distribution
- Most of the indicators for poverty and social exclusion are non linear functions of $y$.

How to estimate correctly their variance? ! ?

In most of the cases : many estimation methods lead to adequate estimator of variance under the commonly used sampling design, no unique methodology!

ATTENTION : Bias may also be a non negligible problem!
Additional problems for Laeken indicators (ex. GINI)

- The distribution of income is hard to model, far away from a (log-)Normal distribution.
- Most of the indicators for poverty and social exclusion are non linear functions of $y$.

How to estimate correctly their variance?!?

In most of the cases: many estimation methods lead to adequate estimator of variance under the commonly used sampling design, no unique methodology!

ATTENTION: Bias may also be a non negligible problem!
Additional problems for Laeken indicators (ex. GINI)

- The distribution of income is hard to model, far away from a (log-)Normal distribution
- Most of the indicators for poverty and social exclusion are non linear functions of $y$.

How to estimate correctly their variance?!?

In most of the cases: many estimation methods lead to adequate estimator of variance under the commonly used sampling design, no unique methodology!

ATTENTION: Bias may also be a non negligible problem!
Additional problems for Laeken indicators (ex. GINI)

- The distribution of income is hard to model, far away from a (log-)Normal distribution
- Most of the indicators for poverty and social exclusion are non linear functions of $y$.

How to estimate correctly their variance?!?

In most of the cases: many estimation methods lead to adequate estimator of variance under the commonly used sampling design, no unique methodology!

ATTENTION: Bias may also be a non negligible problem!
Additional problems for Laeken indicators (ex. GINI)

- The distribution of income is hard to model, far away from a (log-)Normal distribution
- Most of the indicators for poverty and social exclusion are non linear functions of $y$.

How to estimate correctly their variance?!?

In most of the cases: many estimation methods lead to adequate estimator of variance under the commonly used sampling design, no unique methodology!

ATTENTION: Bias may also be a non negligible problem!
How to estimate the variance of complex statistics?

- There are no closed formulae to compute the variance of inequality measures.
- Two main philosophies:
  - **replication methods** (Jacknife, Bootstrap, Balanced repeated replications, Balances Half-Samples, Random Group Methods)
  - **linearization methods** (Taylor Linearization, Linearization based on estimating equations, Linearization based on influence functions, Jacknife Linearization)

Osier (2009), opting for linearization, based on Deville (1999)'s work cleared the way for variance estimation of some main poverty and inequality measures, Graf & Tillé (2013) revisited the results and brought some improvements.
How to estimate the variance of complex statistics?

- There are no closed formulae to compute the variance of inequality measures.
- Two main philosophies:
  - **replication methods** (Jacknife, Bootstrap, Balanced repeated replications, Balances Half-Samples, Random Group Methods)
  - **linearization methods** (Taylor Linearization, Linearization based on estimating equations, Linearization based on influence functions, Jacknife Linearization)

- Osier (2009), opting for linearization, based on Deville (1999)’s work cleared the way for variance estimation of some main poverty and inequality measures, Graf & Tillé (2013) revisited the results and brought some improvements.
How to estimate the variance of complex statistics?

- There are no closed formulae to compute the variance of inequality measures.
- Two main philosophies:
  - **replication methods** (Jacknife, Bootstrap, Balanced repeated replications, Balances Half-Samples, Random Group Methods)
  - **linearization methods** (Taylor Linearization, Linearization based on estimating equations, Linearization based on influence functions, Jacknife Linearization)
- Osier (2009), opting for linearization, based on Deville (1999)’s work cleared the way for variance estimation of some main poverty and inequality measures, Graf & Tillé (2013) revisited the results and brought some improvements.
Indicators for poverty and social exclusion

Let’s focus on five of them, formerly called *Laeken indicators* (Eurostat, 2005)

- At risk of poverty threshold ARPT
- At risk of poverty rate ARPR
- Relative median at-risk-of poverty gap RMPG
- Quintile share ratio \((S_{80}/S_{20})\) QSR
- Gini index GINI
Laeken indicators: empirical calculations

**At risk of poverty threshold** Let \( m \) be the median income,

\[
\text{ARPT} = 0.6m
\]

ARPT is an absolute measure (CHF, EUR, $,...).

**At risk of poverty rate** ARPR is the share of persons with an income \( \leq \text{ARPT} \). Let \( F \) be the distribution function of \( y \),

\[
\text{ARPR} = P(Y < \text{ARPT}) = F^{-1}(\text{ARPT})
\]

ARPR is a relative measure, it’s scale free, \( \text{ARPR} \in [0,1] \).
Laeken indicators : empirical calculations

**At risk of poverty threshold** Let $m$ be the median income,

$$\text{ARPT} = 0.6m$$

ARPT is an absolute measure (CHF, EUR, $,...).

**At risk of poverty rate** ARPR is the share of persons with an income $\leq$ ARPT. Let $F$ be the distribution function of $y$,

$$\text{ARPR} = P(Y < \text{ARPT}) = F^{-1}(\text{ARPT})$$

ARPR is a relative measure, it’s scale free, $\text{ARPR} \in [0, 1]$. 

Laeken indicators: empirical calculations

Relative median at-risk-of poverty gap The RMPG is the relative diff. between the ARPT and the median income of the poors \((m_p)\):

\[
RMPG = \frac{\text{ARPT} - m_p}{\text{ARPT}} \in [0, 1]
\]

Quintile share ratio \((S_{80}/S_{20})\) or degree of income inequality in the population. QSR is the ration of the total income of the persons in the top income quintile \((> q_{80})\) over the total income of the persons in the bottom quintile \((< q_{20})\),

\[
QSR = \frac{\sum y_k > q_{80} y_k}{\sum y_k < q_{20} y_k}
\]
Laeken indicators: empirical calculations

Relative median at-risk-of poverty gap The RMPG is the relative diff. between the ARPT and the median income of the poor ($m_p$):

$$\text{RMPG} = \frac{\text{ARPT} - m_p}{\text{ARPT}} \in [0, 1]$$

Quintile share ratio ($S_{80}/S_{20}$) or degree of income inequality in the population. QSR is the ration of the total income of the persons in the top income quintile ($> q_{80}$) over the total income of the persons in the bottom quintile ($< q_{20}$),

$$\text{QSR} = \frac{\sum_{y_k > q_{80}} y_k}{\sum_{y_k < q_{20}} y_k}$$
Laeken indicators: empirical calculations

**Gini index**

\[
\text{GINI} = 2A \in [0, 1]
\]

\[
A + B = 1/2
\]

\[
G = 2 \int_0^1 [\alpha - L(\alpha)] \, d\alpha.
\]

If the \( y_k \) are \( n \) discrete realization of a random variable \( Y \), sorted by their ranks, an estimation of Gini is given by:

\[
\hat{G} = \frac{2 \sum_{k \in S} ky_k}{n \sum_{k \in S} y_k} - \frac{n + 1}{n}
\]
Knowledge about the income distribution of the population is of vital interest for every study of economic markets in order to govern economic and social decisions.

The study of the distribution of income is in the heart of inequality measures and more generally evaluations of social welfare.

The European project AMELI (2011), relying on the EU-SILC data, showed that generalized beta distribution of the second kind GB2 can be fitted to the income data collected in a good way.
Knowledge about the income distribution of the population is of vital interest for every study of economic markets in order to govern economic and social decisions.

The study of the distribution of income is in the heart of inequality measures and more generally evaluations of social welfare.

The European project AMELI (2011), relying on the EU-SILC data, showed that generalized beta distribution of the second kind GB2 can be fitted to the income data collected in a good way.
Knowledge about the income distribution of the population is of vital interest for every study of economic markets in order to govern economic and social decisions.

The study of the distribution of income is in the heart of inequality measures and more generally evaluations of social welfare.

The European project AMELI (2011), relying on the EU-SILC data, showed that generalized beta distribution of the second kind GB2 can be fitted to the income data collected in a good way.
Generalized beta distribution of the second kind (GB2)

- **four parameters** distribution: \( GB2(a, b, p, q) \).
- has been developed by McDonald (1984).

\[
f_{GB2}(y; a, b, p, q) = \frac{a}{b \cdot B(p, q)} \frac{(y/b)^{ap-1}}{(1+(y/b)^a)^{p+q}}
\]

where \( B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt \) is the beta function.

Empirical studies on income - see for ex. Jenkins (2007); Dastrup et al. (2007); Kleiber et Kotz (2003); Sepanski et Kong (2007) - show that the GB2 fits well with such data and it is often more suitable than other four-parameter distribution.

Results from AMELI (2011) confirm for EU-SILC.
Generalized beta distribution of the second kind (GB2)

- **four parameters** distribution: \( GB2(a, b, p, q) \).
- has been developed by McDonald (1984).

\[
f_{GB2}(y; a, b, p, q) = \frac{a}{b \cdot B(p,q)} \frac{(y/b)^{ap-1}}{(1+(y/b)^a)^{p+q}}
\]

- density: \( B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt \) is the beta function.

Empirical studies on income - see for ex. Jenkins (2007); Dastrup et al. (2007); Kleiber et Kotz (2003); Sepanski et Kong (2007) - show that the GB2 fits well with such data and it is often more suitable than other four-parameter distribution. Results from AMELI (2011) confirm for EU-SILC.
GB2 Law 1

Many other probability distributions can be seen as special cases of the GB2
GB2 Law II

GB2 densities, $a$ varies, $b = 1$, $p = 1$, $q = 0.5$

$\rightarrow$ sharp distribution
GB2 Law III

GB2 densities, $a = 5$, $b = 1$, $p$ varies, $q = 0.5$

→ playing on the shape of the left tail
GB2 Law IV

GB2 densities, \( a = 5, \ b = 1, \ p = 0.25 \), \( q \) varies

→ playing on the shape of the right tail
GB2 Law and inequality measures

Advantage of a parametric estimation for a distribution of income: there are explicit formulas for the inequality measures as functions of the four parameters $\theta = (a, b, p, q)$ of the GB2 adjusted to the data. McDonald (1984), Graf, M. (2009), Ameli (2011)

- At risk of poverty threshold $\text{ARPT}(a,b,p,q)$
- At risk of poverty rate $\text{ARPR}(a,b,p,q)$
- Relative median at-risk-of poverty gap $\text{RMPG}(a,b,p,q)$
- Quintile share ratio ($S_{80}/S_{20}$) $\text{QSR}(a,b,p,q)$
- Gini index $\text{GINI}(a,b,p,q)$

All programmed and available in statistical software R, package("GB2").
GB2 Law and inequality measures

Advantage of a parametric estimation for a distribution of income: there are explicit formulas for the inequality measures as functions of the four parameters $\theta = (a, b, p, q)$ of the GB2 adjusted to the data. McDonald (1984), Graf, M. (2009), Ameli (2011)

- At risk of poverty threshold $\text{ARPT}(a,b,p,q)$
- At risk of poverty rate $\text{ARPR}(a,b,p,q)$
- Relative median at-risk-of poverty gap $\text{RMPG}(a,b,p,q)$
- Quintile share ratio $(S_{80}/S_{20})$ $\text{QSR}(a,b,p,q)$
- Gini index $\text{GINI}(a,b,p,q)$

All programmed and available in statistical software R, package("GB2").
GB2 Law and inequality measures

Advantage of a parametric estimation for a distribution of income: there are explicit formulas for the inequality measures as functions of the four parameters $\theta = (a, b, p, q)$ of the GB2 adjusted to the data. McDonald (1984), Graf, M. (2009), Ameli (2011)

- At risk of poverty threshold $\text{ARPT}(a,b,p,q)$
- At risk of poverty rate $\text{ARPR}(a(b),p,q)$
- Relative median at-risk-of poverty gap $\text{RMPG}(a(b),p,q)$
- Quintile share ratio ($S_{80}/S_{20}$) $\text{QSR}(a(b),p,q)$
- Gini index $\text{GINI}(a(b),p,q)$

All programmed and available in statistical software R, package("GB2").
Generalized linearization technique

- Usable for estimating the variance of sample-based non-linear statistics (Deville, 1999)
- Relies on the concept of influence functions proposed in robust statistic (Hampel, 1974)
- Applied by Osier (2009) for several EU-SILC indicators

Idea: what is the influence of a unit $k$ on the parameter of interest $\theta = T(M)$ viewed as a functional of a measure $M$?

$M$ is a measure allowing some mass to each unit:

- $M(k) = M_k = 1$ for $k \in U$
- $\hat{M}(k) = \hat{M}_k = w_k$ for $k \in U$
Generalized linearization technique

- Usable for estimating the variance of sample-based non linear statistics (Deville, 1999)
- Relies on the concept of influence functions proposed in robust statistic (Hampel, 1974)
- Applied by Osier (2009) for several EU-SILC indicators

Idea: what is the influence of a unit $k$ on the parameter of interest $\theta = T(M)$ viewed as a functional of a measure $M$?

$M$ is a measure allowing some mass to each unit:

- $M(k) = M_k = 1$ for $k \in U$
- $\widehat{M}(k) = \widehat{M}_k = w_k$ for $k \in U$
Influence function

One identifies the influence of the unit $k$ on $\theta$ at the population or sample level by an infinitesimal variation of the weight of the unit.

Influence function of $\theta$

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{ for all } k \in U$$

Empirical influence function of $\theta$

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_k) - T(\hat{M})}{t}, \text{ for all } k \in s$$

Example: for the population total $Y = \sum_{k \in U} y_k$, $Y = T(M)$,

$$\hat{Y} = T(\hat{M}) = \sum_{k \in s} w_k y_k, z_k = y_k, k \in U, \hat{z}_k = w_k y_k, k \in s$$

E. Graf

Variance estimation by linearization
Influence function

One identifies the influence of the unit $k$ on $\theta$ at the population or sample level by an infinitesimal variation of the weight of the unit.

Influence function of $\theta$

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{ for all } k \in U$$

Empirical influence function of $\theta$

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t \to 0} \frac{T(\hat{M} + t\delta_k) - T(\hat{M})}{t}, \text{ for all } k \in s$$

Example: for the population total $Y = \sum_{k \in U} y_k$, $Y = T(M)$, $\hat{Y} = T(\hat{M}) = \sum_{k \in s} w_k y_k$, $z_k = y_k$, $k \in U$, $\hat{z}_k = w_k y_k$, $k \in s$
Generalized linearization technique: main result

Under asymptotic conditions described in Deville (1999), that are in principle satisfied if the sample is "big enough", the variance of the estimated total of $z_k$ is an approximation of the variance of the (complex) statistic $\hat{A}$:

$$\text{var} \left[ \sum_{k \in s} \hat{z}_k w_k(s) \right] \approx \text{var}(\hat{A})$$
Influence functions for poverty and inequality measure


Suppose $y_k$ ordered, $A = \text{GINI}$

$$\hat{z}_k^{\text{GINI}} = \frac{1}{\hat{N}\hat{Y}} \left[ 2\hat{N}_k(y_k - \hat{\bar{y}}_k) + \hat{\bar{y}} - \hat{N}y_k - \hat{G}(\hat{\bar{y}} + y_k\hat{N}) \right]$$

where
- $\hat{N}_k = \sum_{\ell \in s} w_{\ell} 1_{[y_{\ell} \leq y_k]}$ is the cumulated sum of the $w_k$ up to $k$
- $\hat{N} = \sum_{k \in s} w_k$ the estimated total size of the population
- $\hat{\bar{y}}_k = \sum_{\ell \in s} w_{\ell} y_{\ell} 1_{[y_{\ell} \leq y_k]} / \hat{N}_k$ the weighted mean partial sum
Influence functions for poverty and inequality measure


Suppose $y_k$ ordered, $A = \text{GINI}$

\[
\hat{z}_{k}^{\text{GINI}} = \frac{1}{\hat{N}\hat{Y}} \left[ 2\hat{N}_k(y_k - \hat{Y}_k) + \hat{Y} - \hat{N}y_k - \hat{G}(\hat{Y} + y_k\hat{N}) \right]
\]

where

- $\hat{N}_k = \sum_{\ell \leq k} w_\ell \mathbf{1}_{[y_\ell \leq y_k]}$ is the cumulated sum of the $w_k$ up to $k$
- $\hat{N} = \sum_{k \in s} w_k$ the estimated total size of the population
- $\hat{Y}_k = \sum_{\ell \leq s} w_\ell y_\ell \mathbf{1}_{[y_\ell \leq y_k]} / \hat{N}_k$ the weighted mean partial sum
$A = \text{QSR}$ Quintile share ratio

\[
\hat{z}_{QSR}^k = \frac{y_k - \left\{ y_k H \left( \frac{0.8\hat{N}-\hat{N}_k-1}{w_k} \right) + \hat{Q}_{0.8} [0.8 - 1[y_k < \hat{Q}_{0.8}]] \right\}}{\hat{Y}_{0.2}} \\
(\hat{Y} - \hat{Y}_{0.8}) \left\{ y_k H \left( \frac{0.2\hat{N}-\hat{N}_k-1}{w_k} \right) + \hat{Q}_{0.2} [0.2 - 1[y_k < \hat{Q}_{0.2}]] \right\} \\
\hat{Y}_{0.2}^2
\]

where

\[
H(x) = \begin{cases} 
0 & \text{si } x < 0 \\
x & \text{si } 0 \leq x < 1 \\
1 & \text{si } x \geq 1.
\end{cases}
\]
\( A = \text{ARPT} \) At risk of poverty threshold

\[
\hat{z}_{k,\text{ARPT}} = -\frac{0.6}{f(\hat{m}) \frac{1}{N}} \left[ 1[y_k \leq \hat{m}] - 0.5 \right]
\]

\( A = \text{ARPR} \) At risk of poverty rate

\[
\hat{z}_{k,\text{ARPR}} = \frac{1}{N} \left( 1[y_k \leq \text{ARPT}] + \hat{\text{ARPR}} \right) - f(\text{ARPT}) \hat{z}_{k,\text{ARPT}}
\]
$A = \text{ARPT}$ At risk of poverty threshold

$$\hat{z}_{\text{ARPT}} = - \frac{0.6}{f(\hat{m})} \frac{1}{\hat{N}} \left[ 1_{[\hat{y}_k \leq \hat{m}]} - 0.5 \right]$$

$A = \text{ARPR}$ At risk of poverty rate

$$\hat{z}_{\text{ARPR}} = \frac{1}{\hat{N}} \left( 1_{[\hat{y}_k \leq \hat{\text{ARPT}}]} + \hat{\text{ARPR}} \right) - f(\hat{\text{ARPT}}) \hat{z}_{\text{ARPT}}$$
$A = m_p$ Median of poor

$$\hat{z}_{k}^{mp} = \frac{1}{f(\hat{m}_p)} \frac{\hat{z}_{k}^{ARPR}}{2} - \frac{1}{\hat{N}} (1_{[y_k \leq \hat{m}_p]} - F(\hat{m}_p))$$

$A = \text{RMPG}$ Relative median at-risk-of poverty gap

$$\hat{z}_{k}^{\text{RMPG}} = \frac{\hat{m}_p \hat{z}_{k}^{\text{ARPT}} - \widehat{\text{ARPT}} \hat{z}_{k}^{mp}}{\widehat{\text{ARPT}}^2}$$
\( A = m_p \) Median of poor

\[
\hat{z}_{k}^{m_p} = \frac{1}{f(\hat{m}_p)} \frac{\hat{z}_{k}^{ARPR}}{2} - \frac{1}{\hat{N}} \left( 1_{[y_k \leq \hat{m}_p]} - F(\hat{m}_p) \right)
\]

\( A = \text{RMPG} \) Relative median at-risk-of poverty gap

\[
\hat{z}_{k}^{\text{RMPG}} = \frac{\hat{m}_p \hat{z}_{k}^{\text{ARPT}} - \overbrace{\text{ARPT} \hat{z}_{k}^{m_p}}^{\text{ARPT}^2}}{\text{ARPT}^2}
\]
Estimation of income density function \( f \)

The influence functions or linearized variables of the ARPT, ARPR, \( m_p \) and RMPG require to estimate the value of \( f \) in some points. Deville (1999) and Osier (2009) propose to proceed through Gaussian kernel density estimation

\[
\hat{f}_1(x) = \frac{1}{h\sqrt{2\pi}} \frac{1}{N} \sum_{k \in S} w_k \exp \left[ -\frac{(x-y_k)^2}{2h^2} \right]
\]

- \( h \) is the bandwidth that Osier estimates by \( \hat{h} = \hat{\sigma} \hat{N}^{-0.2} \)
- \( \hat{\sigma} \) the empirical standard deviation of the income variable \( Y \).
- The estimation of \( \sigma \) is non robust (sensitive to extreme values)
- In surveys, one encounter often some concentration of observations around some values what may cause trouble with a fixed bandwidth.
Estimation of income density function $f$

The influence functions or linearized variables of the ARPT, ARPR, $m_p$ and RMPG require to estimate the value of $f$ in some points. Deville (1999) and Osier (2009) propose to proceed through Gaussian kernel density estimation

\[ \hat{f}_1(x) = \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \exp\left[ -\frac{(x-y_k)^2}{2h^2} \right] \]

- $h$ is the *bandwidth* that Osier estimates by $\hat{h} = \hat{\sigma} \hat{N}^{-0.2}$
- $\hat{\sigma}$ the empirical standard deviation of the income variable $Y$.
- The estimation of $\sigma$ is non robust (sensitive to extreme values)
- In surveys, one encounter often some *concentration of observations around some values* what may cause trouble with a fixed bandwidth.
Two proposed alternatives for density estimation

2. Estimate through the logarithm

\[ \hat{f}_2(x) = \frac{\hat{f}_1(\log(x))}{x} \]

Diminishes problems due to extreme values in the GK density estimation

3. Nearest neighbours with minimal bandwidth

\[ \hat{f}_{NNMB}(x) = \frac{p(x)}{nh(x)} \]

At each point of the distribution, use at least \( p \) neighbours and a minimal bandwidth \( h(x) \geq h_{opt} \) where \( h_{opt} \) is the rule of thumb of Silverman (1986) to determine the bandwidth. This solution is more robust and avoids problems arising from the concentration of observations around some values.

\[ \hat{f}_3(x) = \frac{\hat{f}_{NNMB}(\log(x))}{x} \]
Two proposed alternatives for density estimation

2. Estimate through the logarithm

\[ \hat{f}_2(x) = \frac{\hat{f}_1(\log(x))}{x} \]

Diminishes problems due to extreme values in the GK density estimation

3. Nearest neighbours with minimal bandwidth

\[ \hat{f}_{\text{NNMB}}(x) = \frac{p(x)}{nh(x)} \]

At each point of the distribution, use at least \( p \) neighbours and a minimal bandwidth \( h(x) \geq h_{opt} \) where \( h_{opt} \) is the rule of thumb of Silverman (1986) to determine the bandwidth. This solution is more robust and avoids problems arising from the concentration of observations around some values.

\[ \hat{f}_3(x) = \frac{\hat{f}_{\text{NNMB}}(\log(x))}{x} \]
Results from simulations

Relative bias of the variance obtained on 1000 simple random samples without replacement on the Swiss SILC 2009 data (y = equivalized household income, N = 17,534)

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Sample sizes (sampling rate)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 500 (2.9%)</td>
<td>n = 1000 (5.7%)</td>
</tr>
<tr>
<td>GINI</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>QSR</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>ARPT</td>
<td>-0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>ARPR</td>
<td>-0.14</td>
<td>-0.10</td>
</tr>
<tr>
<td>RMPG</td>
<td>-0.04</td>
<td>-0.02</td>
</tr>
<tr>
<td>mp</td>
<td>-0.17</td>
<td>-0.15</td>
</tr>
<tr>
<td>m</td>
<td>-0.08</td>
<td>-0.11</td>
</tr>
</tbody>
</table>
## Results from simulations

Relative bias of the variance obtained on 1000 simple random samples without replacement on the Swiss SILC 2009 data ($y = \text{personal income for employed persons, } N = 7,922$)

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Sample sizes (sampling rate)</th>
<th>$n = 500(6.3%)$</th>
<th>$n = 1000(12.6%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{f}_1$</td>
<td>$\hat{f}_2$</td>
<td>$\hat{f}_3$</td>
</tr>
<tr>
<td>GINI</td>
<td>-0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QSR</td>
<td>-0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARPT</td>
<td>0.06</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>ARPR</td>
<td>-0.04</td>
<td>-0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>RMPG</td>
<td>0.53</td>
<td>0.05</td>
<td>-0.03</td>
</tr>
<tr>
<td>$m_p$</td>
<td>0.71</td>
<td>0.14</td>
<td>-0.00</td>
</tr>
<tr>
<td>$m$</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Density estimations for personal income
Variance estim. of indicators under GB2: fitting the GB2


\[ l_n(\theta) = \sum_{i=1}^{n} w_i \log f_{GB2}(y_i; \theta) \]

Partial derivatives of the pseudo-loglikelihood function are set to 0 and solve for \( \theta \):

\[ l'_n(\theta) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \theta} \log f_{GB2}(y_i; \theta) \]

\[ := u_i(y_i; \theta) \]

The scores \( u_i(y_i; \theta) \) are \( 4 \times 1 \) vectors of partial derivatives of the GB2 density function in \( y_i \). They can be seen as the linearized parameters with respect to the GB2.
Variance estim. of indicators under GB2: fitting the GB2


\[ l_n(\theta) = \sum_{i=1}^{n} w_i \log f_{GB2}(y_i; \theta) \]

Partial derivatives of the pseudo-loglikelihood function are set to 0 and solve for \( \theta \):

\[ l'_n(\theta) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \theta} \log f_{GB2}(y_i; \theta) := u_i(y_i; \theta) \]

The scores \( u_i(y_i; \theta) \) are 4×1 vectors of partial derivatives of the GB2 density function in \( y_i \). They can be seen as the linearized parameters with respect to the GB2.
Variance estim. of indicators under GB2 : fitting the GB2

Pseudo maximum likelihood estimation of the parameters of the

\[ l_n(\theta) = \sum_{i=1}^{n} w_i \log f_{GB2}(y_i; \theta) \]

Partial derivatives of the pseudo-loglikelihood function are set to 0
and solve for \( \theta \) :

\[ l'_n(\theta) = \sum_{i=1}^{n} w_i \left( \frac{\partial}{\partial \theta} \log f_{GB2}(y_i; \theta) \right). \]

\[ := u_i(y_i; \theta) \]

The scores \( u_i(y_i; \theta) \) are \( 4 \times 1 \) vectors of partial derivatives of the
GB2 density function in \( y_i \). They can be seen as the linearized
parameters with respect to the GB2.
Density estimations for personal income
Variance estimation of the parameters of the GB2

Approximate \( l'_n(\hat{\theta}_n) \) by the first two terms of a Taylor series around \( \theta \) (linearization!): 

\[
l'_n(\hat{\theta}_n) \approx l'_n(\theta) + l''_n(\theta)(\hat{\theta}_n - \theta) \overset{!}{=} 0
\]

\( \Rightarrow \hat{\theta}_n - \theta \approx [-l''_n(\theta)]^{-1}l'_n(\theta) \)

Get the approximated variance by the so-called Huber sandwich estimator (Huber (1967), White (1980, 1982), Freedman(2006), Pfeffermann and Sverchkov (2003)):

\[
\hat{\text{var}}(\hat{\theta}_n) \approx [l''_n(\hat{\theta}_n)]^{-1}\hat{\text{var}}[l'_n(\hat{\theta})][l''_n(\hat{\theta}_n)]^{-1}
\]

with \( \hat{\text{var}}[l'_n(\hat{\theta}_n)] = \hat{\text{var}}[\sum_{i=1}^{n} w_i u_i(y_i; \hat{\theta}_n)] \) the estimated design variance of the estimated total from the \( u_i \).
Variance estimation of the parameters of the GB2

Approximate $l'_n(\hat{\theta}_n)$ by the first two terms of a Taylor series around $\theta$ (linearization!):

$$
l'_n(\hat{\theta}_n) \approx l'_n(\theta) + l''_n(\theta)(\hat{\theta}_n - \theta) \frac{1}{2} = 0
\Rightarrow \hat{\theta}_n - \theta \approx [-l''_n(\theta)]^{-1}l'_n(\theta)
$$

Get the approximated variance by the so-called Huber sandwich estimator (Huber (1967), White (1980, 1982), Freedman (2006), Pfeffermann and Sverchkov (2003)):

$$\hat{\text{var}}(\hat{\theta}_n) \approx [l''_n(\hat{\theta}_n)]^{-1} \hat{\text{var}}[l'_n(\hat{\theta})][l''_n(\hat{\theta}_n)]^{-1}$$

with $\hat{\text{var}}[l'_n(\hat{\theta}_n)] = \hat{\text{var}}[\sum_{i=1}^n w_i u_i(y_i; \hat{\theta}_n)]$ the estimated design variance of the estimated total from the $u_i$. 

\[\text{(linearization !)}\]
Variance of the indicators expressed as a function of the 4 GB2-parameters \( \theta = (a, b, p, q)' \)

Apply the **Delta method** (Davison (2003), AMELI (2011)).
\( \hat{A} = A(\hat{\theta}_n) \) (for example \( A = \text{GINI, ARPR, QSR,...} \)) then:

\[
\widehat{\text{var}}(\hat{A}) = \frac{\partial \hat{A}'}{\partial \hat{\theta}_n} \widehat{\text{var}}(\hat{\theta}_n) \frac{\partial \hat{A}}{\partial \hat{\theta}_n}
\]

The partial derivatives of the indicators with respect to the 4 parameters can be computed numerically.
Calibrated survey weights and linearization

Weighting scheme of the swiss SILC: main steps (Graf, 2008)

- stratified random cluster sample $\Rightarrow$ sampling weights
- non response corrections on Grid, Household, Individual levels $\Rightarrow$ NR-correction weights
- weight sharing (see Lavallée, 2002)
- panels combinations
- calibration on known totals

Massiani (2012) detailed the procedure to estimate the variance of a total taking all these steps into account, i.e. take the randomness of the weights into account.
Calibrated survey weights and linearization

Weighting scheme of the swiss SILC : main steps (Graf, 2008)

- stratified random cluster sample $\Rightarrow$ sampling weights
- non response corrections on Grid, Household, Individual levels $\Rightarrow$ NR-correction weights
- weight sharing (see Lavallée, 2002)
- panels combinations
- calibration on known totals

Massiani (2012) detailed the procedure to estimate the variance of a total taking all these steps into account, i.e. take the randomness of the weights into account.
Calibrated survey weights and linearization

Weighting scheme of the swiss SILC: main steps (Graf, 2008)

- stratified random cluster sample $\Rightarrow$ sampling weights
- non response corrections on Grid, Household, Individual levels $\Rightarrow$ NR-correction weights
- weight sharing (see Lavallée, 2002)
- panels combinations
- calibration on known totals

Massiani (2012) detailed the procedure to estimate the variance of a total taking all these steps into account, i.e. take the randomness of the weights into account.
Calibrated survey weights and linearization

Weighting scheme of the swiss SILC: main steps (Graf, 2008)

- stratified random cluster sample → sampling weights
- non response corrections on Grid, Household, Individual levels → NR-correction weights
- weight sharing (see Lavallée, 2002)
- panels combinations
- calibration on known totals

Massiani (2012) detailed the procedure to estimate the variance of a total taking all these steps into account, i.e. take the randomness of the weights into account.
Calibrated survey weights and linearization

Weighting scheme of the swiss SILC: main steps (Graf, 2008)

- stratified random cluster sample $\Rightarrow$ sampling weights
- non response corrections on Grid, Household, Individual levels $\Rightarrow$ NR-correction weights
- weight sharing (see Lavallée, 2002)
- panels combinations
- calibration on known totals

Massiani (2012) detailed the procedure to estimate the variance of a total taking all these steps into account, i.e. take the randomness of the weights into account.
Calibrated survey weights and linearization

Weighting scheme of the swiss SILC : main steps (Graf, 2008)

- stratified random cluster sample $\Rightarrow$ sampling weights
- non response corrections on Grid, Household, Individual levels $\Rightarrow$ NR-correction weights
- weight sharing (see Lavallée, 2002)
- panels combinations
- calibration on known totals

Massiani (2012) detailed the procedure to estimate the variance of a total taking all these steps into account, i.e. take the randomness of the weights into account.
Variance estimation for calibrated estimators of totals

- estimated total of $y_k : \sum_{k \in s} w_k y_k$
- estimated total of some linearized variable $z_k : \sum_{k \in s} w_k z_k$
- estimated total of GB2 scores $u_k : \sum_{k \in s} w_k u_k$

Deville (1999) shows that to estimate the variance of such a calibrated estimator (Deville & Särndal, 1992) is equivalent to estimate the design variance of the residuals from the regression of the variable of interest on the auxiliary variables used to calibrate. The residuals of that regression are the linearized variable with respect to the calibration.
Variance estimation for calibrated estimators of totals

- estimated total of $y_k : \sum_{k \in s} w_k y_k$
- estimated total of some linearized variable $z_k : \sum_{k \in s} w_k z_k$
- estimated total of GB2 scores $u_k : \sum_{k \in s} w_k u_k$

Deville (1999) shows that to estimate the variance of such a calibrated estimator (Deville & Särndal, 1992) is equivalent to estimate the design variance of the residuals from the regression of the variable of interest on the auxiliary variables used to calibrate.

$\Rightarrow$ the residuals of that regression are the linearized variable with respect to the calibration.
Variance estimation for calibrated estimators of totals

- estimated total of $y_k : \sum_{k \in s} w_k y_k$
- estimated total of some linearized variable $z_k : \sum_{k \in s} w_k z_k$
- estimated total of GB2 scores $u_k : \sum_{k \in s} w_k u_k$

Deville (1999) shows that to estimate the variance of such a calibrated estimator (Deville & Särndal, 1992) is equivalent to estimate the design variance of the residuals from the regression of the variable of interest on the auxiliary variables used to calibrate. ⇒ the residuals of that regression are the linearized variable with respect to the calibration.
Illustrative results

\[ y = \text{personal income for a subgroup of salaried individuals} \]
\[ n = 7,922 \]
Concluding remarks

- Taylor linearization and generalized linearization techniques are of many use in the filed of survey methodology, these methods work very well with large samples (say > 1000 units).

- In small samples, asymptotic properties may fail and some linearized variables (ex. for the median) can become non robust to extreme values (Croux, 1998)!

- Generalized linearization permits to estimate variances for widely used complex indicators, but one must be careful how the income density is estimated in such computations. We’ve shown some good working methods which reduce the bias of the variance.

- For the poverty and inequality measures considered, empirical and GB2-parametrical calculations provide very similar results.

- Work in progress : take imputation into account in the variance estimation through generalized linearization.
Concluding remarks

- Taylor linearization and generalized linearization techniques are of many use in the filed of survey methodology, these methods work very well with large samples (say > 1000 units).

- In small samples, asymptotic properties may fail and some linearized variables (ex. for the median) can become non robust to extreme values (Croux, 1998).

- Generalized linearization permits to estimate variances for widely used complex indicators, but one must be careful how the income density is estimated in such computations. We’ve shown some good working methods which reduce the bias of the variance.

- For the poverty and inequality measures considered, empirical and GB2-parametrical calculations provide very similar results.

- Work in progress: take imputation into account in the variance estimation through generalized linearization.
Concluding remarks

- Taylor linearization and generalized linearization techniques are of many use in the field of survey methodology, these methods work very well with large samples (say \( > 1000 \) units).

- In small samples, asymptotic properties may fail and some linearized variables (e.g., for the median) can become non-robust to extreme values (Croux, 1998).

- Generalized linearization permits to estimate variances for widely used complex indicators, but one must be careful how the income density is estimated in such computations. We've shown some good working methods which reduce the bias of the variance.

- For the poverty and inequality measures considered, empirical and GB2-parametrical calculations provide very similar results.

- Work in progress: take imputation into account in the variance estimation through generalized linearization.
Concluding remarks

- Taylor linearization and generalized linearization techniques are of many use in the field of survey methodology, these methods work very well with large samples (say > 1000 units).

- In small samples, asymptotic properties may fail and some linearized variables (e.g., for the median) can become non-robust to extreme values (Croux, 1998).

- Generalized linearization permits to estimate variances for widely used complex indicators, but one must be careful how the income density is estimated in such computations. We’ve shown some good working methods which reduce the bias of the variance.

- For the poverty and inequality measures considered, empirical and GB2-parametrical calculations provide very similar results.

- Work in progress: take imputation into account in the variance estimation through generalized linearization.
Concluding remarks

- Taylor linearization and generalized linearization techniques are of many use in the field of survey methodology, these methods work very well with large samples (say $> 1000$ units).

- In small samples, asymptotic properties may fail and some linearized variables (ex. for the median) can become non robust to extreme values (Croux, 1998)!

- Generalized linearization permits to estimate variances for widely used complex indicators, but one must be careful how the income density is estimated in such computations. We’ve shown some good working methods which reduce the bias of the variance.

- For the poverty and inequality measures considered, empirical and GB2-parametrical calculations provide very similar results.

- Work in progress: take imputation into account in the variance estimation through generalized linearization.
Cited references

- Eurostat (2005). Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. *The continuity of indicators during the transition between ECHP and EU-SILC.*


European projects : DACSEIS, KEI, SAMPLE, AMELI, see [http://www.uni-trier.de](http://www.uni-trier.de)