Poverty and Inequality with Complex Survey Data

Guilherme Jacob, Anthony Damico, and Djalma Pessoa

2023-08-29

Contents

1 Introduction **5**

1.1 Installation 5

1.2 Complex surveys and statistical inference 6

1.3 Usage Examples 6

1.4 Underlying Calculations 10

1.5 The Variance Estimator 11

1.6 Influence Functions 12

1.7 Influence Function Examples 13

1.8 Examples of Linearization Using the Influence Function 13

1.9 Replication Designs 14

1.10 Decomposition 14

2 Poverty Indices **15**

2.1 At Risk of Poverty Threshold (svyarpt) 15

2.2 At Risk of Poverty Ratio (svyarpr) 20

2.3 Relative Median Income Ratio (svyrmir) 28

2.4 Relative Median Poverty Gap (svyrmpg) 32

2.5 Median Income Below the At Risk of Poverty Threshold (svypoormed) 36

2.6 Foster-Greer-Thorbecke class (svyfgt, svyfgtdec) 41

2.7 Watts poverty measure (svywatts, svywattsdec) 52

3 Inequality Measurement **63**

3.1 The Gender Pay Gap (svygpg) 63
3.2 Quintile Share Ratio (svyqsr) 68
3.3 Lorenz Curve (svylorenz) 73
3.4 Gini index (svygini) 76
3.5 Zenga index (svyzenga) 81
3.6 Entropy-based Measures 86
3.7 Generalized Entropy and Decomposition (svygei, svygeidec) . . . 87
3.8 J-Divergence and Decomposition (svyjdiv, svyjdivdec) 94
3.9 Atkinson index (svyatk) 100
3.10 Which inequality measure should be used? 104

4 Richness Measures **107**

Chapter 1

Introduction

The R `convey` library estimates measures of poverty, inequality, and wellbeing. There are two other R libraries covering this subject, vardpoor (Breidaks et al., 2020) and laeken (Alfons et al., 2014), however, only `convey` integrates seamlessly with the R survey package (Lumley, 2004, 2010, 2020).

`convey` is free and open-source software that runs inside the R environment for statistical computing. Anyone can review and propose changes to the source code for this software. Readers are welcome to propose changes to this book as well.

Individuals getting started in the field of poverty and inequality statistics might find the number of techniques described in this textbook overwhelming, especially on the topic of choosing which method might be most appropriate for each particular research question. For this reason, the authors of this textbook consider Dr. Ija Trapeznikova's article Measuring income inequality an important summary of how to approach selecting between available techniques.

1.1 Installation

In order to work with the `convey` library, you will need to have R running on your machine. If you have never used R before, you will need to install that software before `convey` can be accessed. Once you have R loaded on your machine, you can install..

- the latest released version from CRAN with

```
install.packages("convey")
```

- the latest development version from github with

```
remotes::install_github("ajdamico/convey")
```

In order to know how to cite this package, run `citation("convey")`.

1.2 Complex surveys and statistical inference

In this book, we demonstrate how to measure poverty and income concentration in a population based on microdata collected from a complex survey sample. Most surveys administered by government agencies or larger research organizations utilize a sampling design that violates the assumption of simple random sampling (SRS), including:

1. Different units selection probabilities;
2. Clustering of units;
3. Stratification of clusters;
4. Reweighting to compensate for missing values and other adjustments.

Therefore, basic unweighted R commands such as `mean()` or `glm()` will not properly account for the weighting nor the measures of uncertainty (such as the confidence intervals) present in the dataset. For some examples of publicly-available complex survey data sets, see http://asdfree.com.

Unlike other software, the R `convey` package does not require that the user specify these parameters throughout the analysis. So long as the svydesign object or svrepdesign object has been constructed properly at the outset of the analysis, the `convey` package will incorporate the survey design automatically and produce statistics and variances that take the complex sample into account.

Survey analysts familiar with the R `dplyr` syntax implemented by the `survey` library's wrapper `srvyr` package might be interested in implementing specific `convey` functions by following the `svygini()` example published by `srvyr` author Greg Freedman Ellis. Note that the full design stored by `convey_prep()` may in some cases complicate this extension.

1.3 Usage Examples

In the following example, we've loaded the data set `eusilc` from the R library laeken (Alfons et al., 2014).

```
library(laeken)
data(eusilc)
```

Next, we create an object of class `survey.design` using the function `svydesign` of the `survey` library:

```
library(survey)
des_eusilc <-
  svydesign(
    ids = ~ rb030,
    strata =  ~ db040,
    weights = ~ rb050,
    data = eusilc
  )
```

Right after the creation of the design object `des_eusilc`, we should use the function `convey_prep` that adds an attribute to the survey design which saves information on the design object based upon the whole sample, needed to work with subset designs.

```
library(convey)
des_eusilc <- convey_prep(des_eusilc)
```

To estimate the at-risk-of-poverty rate, we use the function `svyarpt`:

```
svyarpr( ~ eqIncome, design = des_eusilc)
```

```
            arpr     SE
eqIncome 0.14444 0.0028
```

To estimate the at-risk-of-poverty rate across domains defined by the variable `db040` we use:

```
svyby(
  ~ eqIncome,
  by = ~ db040,
  design = des_eusilc,
  FUN = svyarpr,
  deff = FALSE
)
```

```
                      db040  eqIncome          se
Burgenland       Burgenland 0.1953984 0.017202852
Carinthia         Carinthia 0.1308627 0.010606502
Lower Austria Lower Austria 0.1384362 0.006513217
Salzburg           Salzburg 0.1378734 0.011581408
```

```
Styria               Styria 0.1437464 0.007453192
Tyrol                 Tyrol 0.1530819 0.009884094
Upper Austria Upper Austria 0.1088977 0.005933094
Vienna               Vienna 0.1723468 0.007684540
Vorarlberg       Vorarlberg 0.1653731 0.013756389
```

Using the same data set, we estimate the quintile share ratio:

```
# for the whole population
svyqsr( ~ eqIncome, design = des_eusilc, alpha1 = .20)
```

```
          qsr     SE
eqIncome 3.97 0.0426
```

```
# for domains
svyby(
  ~ eqIncome,
  by = ~ db040,
  design = des_eusilc,
  FUN = svyqsr,
  alpha1 = .20,
  deff = FALSE
)
```

```
                     db040 eqIncome         se
Burgenland       Burgenland 5.008486 0.32755685
Carinthia         Carinthia 3.562404 0.10909726
Lower Austria Lower Austria 3.824539 0.08783599
Salzburg           Salzburg 3.768393 0.17015086
Styria               Styria 3.464305 0.09364800
Tyrol                 Tyrol 3.586046 0.13629739
Upper Austria Upper Austria 3.668289 0.09310624
Vienna               Vienna 4.654743 0.13135731
Vorarlberg       Vorarlberg 4.366511 0.20532075
```

These functions can be used as S3 methods for the classes `survey.design` and `svyrep.design`.

Let's create a design object of class `svyrep.design` and run the function `convey_prep` on it:

```
des_eusilc_rep <- as.svrepdesign(des_eusilc, type = "bootstrap")
des_eusilc_rep <- convey_prep(des_eusilc_rep)
```

The function `svyarpr` produces matching coefficients and near-identical standard errors on the replication design:

```
svyarpr( ~ eqIncome, design = des_eusilc_rep)
```

```
            arpr     SE
eqIncome 0.14444 0.0025
```

```
svyby(
  ~ eqIncome,
  by = ~ db040,
  design = des_eusilc_rep,
  FUN = svyarpr,
  deff = FALSE
)
```

```
                      db040  eqIncome se.eqIncome
Burgenland       Burgenland 0.1953984 0.016713791
Carinthia         Carinthia 0.1308627 0.012061625
Lower Austria Lower Austria 0.1384362 0.007294696
Salzburg           Salzburg 0.1378734 0.010050357
Styria               Styria 0.1437464 0.008558783
Tyrol                 Tyrol 0.1530819 0.010328225
Upper Austria Upper Austria 0.1088977 0.006212301
Vienna               Vienna 0.1723468 0.007259732
Vorarlberg       Vorarlberg 0.1653731 0.012792618
```

The functions of the convey `library` are called in a similar way to the functions in `survey` library.

It is also possible to deal with missing values by using the argument `na.rm`.

```
# survey.design using a variable with missings
svygini( ~ py010n , design = des_eusilc)
```

```
       gini SE
py010n   NA NA
```

```
svygini( ~ py010n , design = des_eusilc , na.rm = TRUE)
```

```
          gini     SE
py010n 0.64606 0.0036
```

```
# svyrep.design using a variable with missings
svygini( ~ py010n , design = des_eusilc_rep)
```

```
       gini SE
py010n   NA NA
```

```
svygini( ~ py010n , design = des_eusilc_rep , na.rm = TRUE)
```

```
          gini     SE
py010n 0.64606 0.0043
```

1.4 Underlying Calculations

In what follows, we often use the linearization method as a tool to produce an approximation for the variance of an estimator. From the linearized variable z of an estimator T, we get from the expression 1.5 an estimate of the variance of T.

If T can be expressed as a function of the population totals $T = g(Y_1, Y_2, \dots, Y_n)$, and if g is linear, the estimation of the variance of $T = g(Y_1, Y_2, \dots, Y_n)$ is straightforward. If g is not linear but is a 'smooth' function, then it is possible to approximate the variance of $g(Y_1, Y_2, \dots, Y_n)$ by the variance of its first order Taylor expansion. For example, we can use Taylor expansion to linearize the ratio of two totals. However, there are situations where Taylor linearization cannot be immediately computed, either because T cannot be expressed as functions of the population totals, or because g is not a `smooth` function. A common example is the case where T is a quantile.

In these cases, an alternative form of linearization of T might suffice: the `Influence Function`, as defined in 1.4, proposed in Deville (1999). Separately, replication methods such as `bootstrap` and `jackknife` also work.

In the `convey` library, there are some basic functions that produce the linearized variables needed to measure income concentration and poverty. For example, looking at the income variable in some complex survey dataset, the `quantile` of that income variable can be linearized by the function `convey::svyiqalpha` and the sum total below any quantile of the variable is linearized by the function `convey::svyisq`.

From the linearized variables of these basic estimates, it is possible by using rules of composition, valid for influence functions, to derive the influence function of more complex estimates. By definition, the influence function is a Gateaux derivative and the rules of composition valid for Gateaux derivatives also hold for Influence Functions.

The following property of Gateaux derivatives is commonly used in the `convey` library: Let g be a differentiable function of m variables. Suppose we want to compute the influence function of the estimator $g(T_1, T_2, ..., T_m)$, knowing the influence function of the estimators $T_i, i = 1, ..., m$. Then the following holds:

$$I(g(T_1, T_2, ..., T_m)) = \sum_{i=1}^{m} \frac{\partial g}{\partial T_i} I(T_i)$$

In the `convey` library, this rule is implemented by the function `contrastinf`, which uses the base R function `deriv` to compute the formal partial derivatives $\frac{\partial g}{\partial T_i}$.

For example, suppose we want to linearize the relative median poverty gap (RMPG), defined as the difference between the at-risk-of-poverty threshold (ARPT) and the median of incomes less than the ARPT, relative to the ARPT itself. Let's say that this median income below the at-risk-of-poverty-threshold (POORMED) is the median of incomes less than ARPT:

$$rmpg = \frac{(arpt - poormed)}{arpt}$$

If we know how to linearize ARPT and POORMED, then by applying the function `contrastinf` with

$$g(T_1, T_2) = \frac{(T_1 - T_2)}{T_1}$$

we are also able to linearize the RMPG.

1.5 The Variance Estimator

Using the notation in Osier (2009), the variance of the estimator $T(\hat{M})$ can approximated by:

$$Var\left[T(\hat{M})\right] \cong var\left[\sum_s w_i z_i\right] (\#var) \tag{1.1}$$

The `linearized` variable z is given by the derivative of the functional:

$$z_k = lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t} = IT_k(M)(\#lin) \tag{1.2}$$

where, δ_k is the Dirac measure in k: $\delta_k(i) = 1$ if and only if $i = k$.

This **derivative** is called the **Influence Function** and was introduced in the area of **Robust Statistics**.

1.6 Influence Functions

Some measures of poverty and income concentration are defined by non-differentiable functions, so that it is not always possible to use Taylor linearization to estimate variances. An alternative is to use **influence functions** as described in Deville (1999) and Osier (2009). The `convey` library implements this methodology to work with `survey.design` objects and also with `svyrep.design` objects.

Some examples of these measures are:

- At-risk-of-poverty threshold: $arpt = .60q_{.50}$ where $q_{.50}$ is the median income;
- At-risk-of-poverty rate: $arpr = \frac{\sum_U 1(y_i \leq arpt)}{N}.100$
- Quintile share ratio: $qsr = \frac{\sum_U 1(y_i > q_{.80})}{\sum_U 1(y_i \leq q_{.20})}$
- Gini coefficient $1 + G = \frac{2\sum_U (r_i - 1)y_i}{N\sum_U y_i}$ where r_i is the rank of y_i.

Note that it is not possible to use Taylor linearization for these measures because they rely on quantiles or, in the case of the Gini coefficient, a function of ranks. Therefore, we instead follow the approach proposed by Deville (1999) based upon influence functions.

Let U be a population of size N and M be a measure that allocates mass one to the set composed by one unit, that is $M(i) = M_i = 1$ if $i \in U$ and $M(i) = 0$ if $i \notin U$.

Now, a population parameter θ can be expressed as a functional of M $\theta = T(M)$.

Examples of such parameters are:

- Total: $Y = \sum_U y_i = \sum_U y_i M_i = \int y dM = T(M)$
- Ratio of two totals: $R = \frac{Y}{X} = \frac{\int y dM}{\int x dM} = T(M)$
- Cumulative distribution function: $F(x) = \frac{\sum_U 1(y_i \leq x)}{N} = \frac{\int 1(y \leq x) dM}{\int dM} = T(M)$

To estimate these parameters from the sample, we replace the measure M by the estimated measure $\hat{M}$ defined by: $\hat{M}(i) = \hat{M}_i = w_i$ if $i \in s$ and $\hat{M}(i) = 0$ if $i \notin s$.

The estimators of the population parameters can then be expressed as functional of the measure $\hat{M}$.

- Total: $\hat{Y} = T(\hat{M}) = \int y d\hat{M} = \sum_s w_i y_i$
- Ratio of totals: $\hat{R} = T(\hat{M}) = \frac{\int y d\hat{M}}{\int x d\hat{M}} = \frac{\sum_s w_i y_i}{\sum_s w_i x_i}$
- Cumulative distribution function: $\hat{F}(x) = T(\hat{M}) = \frac{\int 1(y \leq x) d\hat{M}}{\int d\hat{M}} = \frac{\sum_s w_i 1(y_i \leq x)}{\sum_s w_i}$

1.7 Influence Function Examples

- Total:

$$\begin{aligned} IT_k(M) &= lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t} \\ &= lim_{t \to 0} \frac{\int y.d(M + t\delta_k) - \int y.dM}{t} \\ &= lim_{t \to 0} \frac{\int y d(t\delta_k)}{t} = y_k \end{aligned}$$

- Ratio of two totals:

$$\begin{aligned} IR_k(M) &= I\left(\frac{U}{V}\right)_k (M) = \frac{V(M) \times IU_k(M) - U(M) \times IV_k(M)}{V(M)^2} \\ &= \frac{Xy_k - Yx_k}{X^2} = \frac{1}{X}(y_k - Rx_k) \end{aligned}$$

1.8 Examples of Linearization Using the Influence Function

- At-risk-of-poverty threshold:

$$arpt = 0.6 \times m$$

where m is the median income.

$$z_k = -\frac{0.6}{f(m)} \times \frac{1}{N} \times [I(y_k \leq m - 0.5)]$$

- At-risk-of-poverty rate:

$$arpr = \frac{\sum_U I(y_i \leq t)}{\sum_U w_i}.100$$

$$z_k = \frac{1}{N}\left[I(y_k \leq t) - t\right] - \frac{0.6}{N} \times \frac{f(t)}{f(m)}\left[I(y_k \leq m) - 0.5\right]$$

where:

N - population size;

t - at-risk-of-poverty threshold;

y_k - income of person k;

m - median income;

f - income density function;

1.9 Replication Designs

All major functions in the `convey` library have S3 methods for the classes: `survey.design`, `svyrep.design` and `DBIdesign`. When the argument `design` is a survey design with replicate weights created by the `survey` library, `convey` uses the method `svrepdesign`.

Considering the remarks in (Wolter, 1985), p. 163, concerning the deficiency of the `Jackknife` method in estimating the variance of `quantiles`, we adopted the bootstrap method instead.

The function `bootVar` from the `laeken` library (Alfons et al., 2014), also uses the bootstrap method to estimate variances.

1.10 Decomposition

Some inequality and multidimensional poverty measures can be decomposed. The decomposition methods in the `convey` library are limited to group decomposition.

For instance, the generalized entropy index can be decomposed into between and within group components. This sheds light on a very simple question: *of the overall inequality, how much can be explained by inequalities between groups and within groups?* Since this measure is additively decomposable, one can get estimates of the coefficients, SEs and covariance between components. For a more practical approach, see (Lima, 2013).

The Alkire-Foster class of multidimensional poverty indices (not implemented in the `convey` library) can be decomposed by both dimensions and by groups, showing how much each group and each dimension contributes to poverty overall. Multidimensional poverty techniques can sometimes help economic and policy analysts understand more precisely who is more affected by inequality and poverty, and how those disparities manifest.

Chapter 2

Poverty Indices

Poverty has been a topic of conversation throughout human history. As Ravallion (2016) points out, Aristotle and Confucius discussed ideas about poverty. In fact, Aristotle's ideas influenced Thomas Aquinas, one of the pillars of Western philosophy. Since then, societies changed, modifying the theories of justice underlying the idea of poverty.

As the concept and the ethics towards poverty change, so does its measurement. From basic measures like the headcount rate to more complex metrics, such as the Foster-Greer-Thorbecke (FGT) index, poverty measurement has evolved. Nowadays, poverty estimates are calculated using household surveys and censuses (Deaton, 1997). However, only recently have the aspects of statistical inference combining such measures and complex sample survey designs been explored (for more discussion on this topic, see Deville (1999), Berger and Skinner (2003), Bhattacharya (2007), and Osier (2009)). These advances became even more important given the recent efforts in poverty mapping, an analytical method that combines poverty analysis and small area estimation, like Elbers et al. (2003) and Bedi et al. (2007).

This chapter shows how poverty estimates and their sampling errors can be estimated using simple commands from the `convey` package.

2.1 At Risk of Poverty Threshold (svyarpt)

The at-risk-of-poverty threshold (ARPT) is a measure used to define the people whose incomes imply a low standard of living in comparison to the general living standards. Even though some people are not below the effective poverty line, those below the ARPT can be considered "almost deprived".

This measure is defined as 0.6 times the median income for the entire population:

$$arpt = 0.6 \times median(y),$$

where y is the income variable and `median` is estimated for the whole population. The details of the linearization of the ARPT are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes an ARPT coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the laeken library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the arpt coefficient
# using the R vardpoor library
varpoord_arpt_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
```

```
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # arpt coefficient function
    type = "linarpt",

    # get linearized variable
    outp_lin = TRUE
  )


# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )

# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_arpt_calculation$all_result$value
```

```
## [1] 10859.24
```

```
coef(svyarpt( ~ eqincome , des_eusilc))
```

```
## eqincome
## 10859.24
```

```
# linearized variables do match
# vardpoor
lin_arpt_varpoord <- varpoord_arpt_calculation$lin_out$lin_arpt
# convey
lin_arpt_convey <- attr(svyarpt( ~ eqincome , des_eusilc), "lin")

# check equality
all.equal(lin_arpt_varpoord, lin_arpt_convey)
```

```
## [1] TRUE
```

```
# variances do not match exactly
attr(svyarpt( ~ eqincome , des_eusilc) , 'var')
```

```
##          eqincome
## eqincome 2564.027
```

```
varpoord_arpt_calculation$all_result$var
```

```
## [1] 2559.442
```

```
# standard errors do not match exactly
varpoord_arpt_calculation$all_result$se
```

```
## [1] 50.59093
```

```
SE(svyarpt( ~ eqincome , des_eusilc))
```

```
##          eqincome
## eqincome 50.63622
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svyarpt` and `vardpoor::linarpt` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )

# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)



# matches
stopifnot(all.equal(
  attr(svyarpt( ~ eqincome , des_eusilc_ultimate_cluster) , 'var')[1] ,
  varpoord_arpt_calculation$all_result$var
))

# matches
stopifnot(all.equal(varpoord_arpt_calculation$all_result$se ,
                    SE(
                      svyarpt( ~ eqincome , des_eusilc_ultimate_cluster)
                    )[1]))
```

For additional usage examples of `svyarpt`, type `?convey::svyarpt` in the R console.

2.2 At Risk of Poverty Ratio (svyarpr)

The at-risk-of-poverty rate (ARPR) is the share of persons with an income below the at-risk-of-poverty threshold (ARPT). The logic behind this measure is that although most people below the ARPT cannot be considered "poor", they are the ones most vulnerable to becoming poor in the event of a negative economic phenomenon like a recession.

The ARPR is a composite estimate, taking into account both the sampling error in the proportion itself and that in the ARPT estimate. The details of the linearization of the ARPR are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a ARPR coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the vardpoor library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the arpr coefficient
# using the R vardpoor library
```

```
varpoord_arpr_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # arpr coefficient function
    type = "linarpr",

    # get linearized variable
    outp_lin = TRUE

  )


# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )

# immediately run the convey_prep function on it
```

```
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_arpr_calculation$all_result$value
```

```
## [1] 14.44422
```

```
coef(svyarpr( ~ eqincome , des_eusilc)) * 100
```

```
## eqincome
## 14.44422
```

```
# linearized variables do not match
# because Fprime is the derivative wrt
# to the estimated threshold, not the estimated quantile
# for more details, see
# https://github.com/ajdamico/convey/issues/372#issuecomment-1656264143
#
# vardpoor
lin_arpr_varpoord <- varpoord_arpr_calculation$lin_out$lin_arpr
# convey
lin_arpr_convey <- attr(svyarpr( ~ eqincome , des_eusilc), "lin")

# check equality
all.equal(lin_arpr_varpoord, 100 * lin_arpr_convey)
```

```
## [1] "Mean relative difference: 0.2264738"
```

```
# variances do not match exactly
attr(svyarpr( ~ eqincome , des_eusilc) , 'var') * 10000
```

```
##            eqincome
## eqincome 0.07599778
```

```
varpoord_arpr_calculation$all_result$var
```

```
## [1] 0.08718569
```

```
# standard errors do not match exactly
varpoord_arpr_calculation$all_result$se
```

```
## [1] 0.2952722
```

```
SE(svyarpr( ~ eqincome , des_eusilc)) * 100
```

```
##           eqincome
## eqincome 0.2756769
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svyarpr` and `vardpoor::linarpr` produce the same linearized variable z.

However, the measures of uncertainty do not line up. One of the reasons is that `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )

# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)

# does not match
attr(svyarpr( ~ eqincome , des_eusilc_ultimate_cluster) , 'var') * 10000
```

```
##            eqincome
## eqincome 0.07586194
```

```
varpoord_arpr_calculation$all_result$var
```

```
## [1] 0.08718569
```

```
# does not match
varpoord_arpr_calculation$all_result$se
```

```
## [1] 0.2952722
```

```
SE(svyarpr( ~ eqincome , des_eusilc_ultimate_cluster)) * 100
```

```
##           eqincome
## eqincome 0.2754305
```

Still, there is a difference in the estimates. This is discussed in detail in this issue. In order to still provide additional examples for our code, we proceed with a Monte Carlo experiment.

Using the `eusilcP` data from the `simPop` package (Templ et al., 2017), we can compute the actual value of the at risk of poverty rate for that population:

```
# load libraries
library(sampling)
library(survey)
library(convey)
library(parallel)

# load pseudo population data
data("eusilcP" , package = "simPop")

# compute population median
q50.pop <-
  convey:::computeQuantiles(eusilcP$eqIncome , rep(1 , length(eusilcP$eqIncome)) , .50)

# compute population poverty threshold
# as 60% of the median
thresh.pop <- .60 * q50.pop

# compute population at risk of poverty rate
(theta.pop <-
    mean(eusilcP$eqIncome <= thresh.pop , na.rm = TRUE))
```

```
## [1] 0.1469295
```

Now, to study the distribution of the estimator under a particular sampling design, we select 5000 samples under one-stage cluster sampling of 100 households using the `cluster` function from the `sampling` package (Tillé & Matei, 2021), and use the `svyarpr` function to estimate the ARPR for each of those samples:

```
# define the number of monte carlo replicates
mc.rep <- 5000L

# simulation function
arpr_sim_fun <- function(this.iter) {

  set.seed(this.iter)


  library(survey)
  library(convey)
  library(sampling)

  # load pseudo population data
  data("eusilcP" , package = "simPop")

  # compute size-like variable for PPS sampling design
  eusilcP$aux <-
    log(ifelse(eusilcP$eqIncome >= 1000 , eusilcP$eqIncome , 1000))


  # select sample
  tt <-
    sampling::cluster(
      data = eusilcP[sample.int(nrow(eusilcP) , nrow(eusilcP) , replace = FALSE) , ] ,
      clustername = "hid" ,
      size = 1000L ,
      method = "systematic" ,
      pik = eusilcP$aux
    )

  # collect data
  this.sample <- getdata(eusilcP , tt)

  # create survey design object
  this.desobj <-
    svydesign(
      ids = ~ hid ,
      probs = ~ Prob ,
      data = this.sample ,
      nest = FALSE
```

```
    )

  # prepare for convey functions
  this.desobj <- convey_prep(this.desobj)

  # compute estimates
  svyarpr( ~ eqIncome , this.desobj)

}

# run replications
cl <- makeCluster(detectCores() - 1)

arpr.estimate.list <-
  clusterApply(cl, seq_len(mc.rep) , arpr_sim_fun)

stopCluster(cl)
```

Then, we evaluate the Percentage Relative Bias (PRB) of the ARPR estimator. Under this scenario, the PRB of the ARPR estimator is -0.19830%.

```
# estimate the expected value of the ARPR estimator
# using the average of the estimates
(theta.exp <- mean(sapply(arpr.estimate.list , coef)))
```

```
## [1] 0.1466381
```

```
# estimate the percentage relative bias
(percentage_relative_bias_arpr <- 100 * (theta.exp / theta.pop - 1))
```

```
## [1] -0.1982981
```

```
stopifnot(round(percentage_relative_bias_arpr , 4) == -0.19830)
```

For the variance estimator, we have:

```
# estimate the variance of the ARPR estimator
# using the empirical variance of the estimates
(vartheta.popest <- var(sapply(arpr.estimate.list , coef)))
```

```
## [1] 0.0001048728
```

```
# estimate the expected value of the ARPR variance estimator
# using the (estimated) expected value of the variance estimates
(vartheta.exp <- mean(sapply(arpr.estimate.list , vcov)))
```

```
## [1] 0.0001081994
```

```
# estimate the percentage relative bias of the variance estimator
( percentage_relative_bias_variance <- 100 *  (vartheta.exp / vartheta.popest - 1) )
```

```
## [1] 3.172023
```

```
stopifnot( round( percentage_relative_bias_variance , 4 ) == 3.1720 )
```

Under this scenario, the PRB of the ARPR variance estimator is 3.1720%.

Our simulations show that the Squared Bias of this estimator accounts for less than 0.1% of its Mean Squared Error:

```
theta.bias2 <- (theta.exp - theta.pop) ^ 2
theta.mse <- theta.bias2 + vartheta.popest
(squared_bias_over_mse <-  100 * (theta.bias2 / theta.mse))
```

```
## [1] 0.04103177
```

```
stopifnot(squared_bias_over_mse < 0.1)
```

Next, we evaluate the Percentage Coverage Rate (PCR). In theory, under repeated sampling, the estimated 95% CIs should cover the population parameter approximately 95% of the time. We can evaluate that using:

```
# estimate confidence intervals of the ARPR
# for each of the samples
est.coverage <-
  sapply(arpr.estimate.list, function(this.stat)
    confint(this.stat)[, 1] <= theta.pop &
      confint(this.stat)[, 2] >= theta.pop)

# evaluate empirical coverage
(empirical_coverage <- mean(est.coverage))
```

```
## [1] 0.9516
```

```
stopifnot(round(empirical_coverage , 2) == 0.95)
```

Our coverages are not too far from the nominal coverage level of 95%.

For additional usage examples of `svyarpr`, type `?convey::svyarpr` in the R console.

2.3 Relative Median Income Ratio (svyrmir)

The relative median income ratio (RMIR) is the ratio of the median income of people aged above a value (65) to the median of people aged below the same value. In mathematical terms,

$$rmir = \frac{median\{y_i; age_i > 65\}}{median\{y_i; age_i \leq 65\}}.$$

The details of the linearization of the RMIR and are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a RMIR coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the vardpoor library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)
```

```r
# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the rmir coefficient
# using the R vardpoor library
varpoord_rmir_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # age variable
    age = "age",

    # rmir coefficient function
    type = "linrmir",

    # get linearized variable
    outp_lin = TRUE

  )
```

```
# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )

# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_rmir_calculation$all_result$value
```

```
## [1] 0.9330361
```

```
coef(svyrmir( ~ eqincome , des_eusilc, age = ~ age))
```

```
##  eqincome
## 0.9330361
```

```
# linearized variables do match
# vardpoor
lin_rmir_varpoord <- varpoord_rmir_calculation$lin_out$lin_rmir
# convey
lin_rmir_convey <-
  attr(svyrmir( ~ eqincome , des_eusilc, age = ~ age), "lin")

# check equality
all.equal(lin_rmir_varpoord, lin_rmir_convey[, 1])
```

```
## [1] TRUE
```

```
# variances do not match exactly
attr(svyrmir( ~ eqincome , des_eusilc, age = ~ age) , 'var')
```

```
##             eqincome
## eqincome 0.000127444
```

```
varpoord_rmir_calculation$all_result$var
```

```
## [1] 0.0001272137
```

```
# standard errors do not match exactly
varpoord_rmir_calculation$all_result$se
```

```
## [1] 0.0112789
```

```
SE(svyrmir( ~ eqincome , des_eusilc , age = ~ age))
```

```
##            eqincome
## eqincome 0.01128911
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svyrmir` and `vardpoor::linrmir` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )
```

```
# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)

# matches
stopifnot(all.equal(
  attr(
    svyrmir(~ eqincome , des_eusilc_ultimate_cluster , age = ~ age) ,
    'var'
  )[1],
  varpoord_rmir_calculation$all_result$var
))


# matches
stopifnot(all.equal(SE(
  svyrmir(~ eqincome , des_eusilc_ultimate_cluster, age = ~ age)
)[1], varpoord_rmir_calculation$all_result$se))
```

For additional usage examples of `svyrmir`, type `?convey::svyrmir` in the R console.

2.4 Relative Median Poverty Gap (svyrmpg)

The relative median poverty gap (RMPG) is the relative difference between the median income of people having income below the ARPT and the ARPT itself:

$$rmpg = \frac{median\{y_i, y_i < arpt\} - arpt}{arpt}$$

The details of the linearization of the RMPG are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a RMPG coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```r
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the vardpoor library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the rmpg coefficient
# using the R vardpoor library
varpoord_rmpg_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",
```

```
    # data.table
    dataset = dati,

    # rmpg coefficient function
    type = "linrmpg",

    # get linearized variable
    outp_lin = TRUE

  )



# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )

# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_rmpg_calculation$all_result$value
```

```
## [1] 18.9286
```

```
coef(svyrmpg( ~ eqincome , des_eusilc)) * 100
```

```
## eqincome
##  18.9286
```

```
# linearized variables do match
# vardpoor
lin_rmpg_varpoord <- varpoord_rmpg_calculation$lin_out$lin_rmpg
# convey
lin_rmpg_convey <- attr(svyrmpg( ~ eqincome , des_eusilc), "lin")

# check equality
all.equal(lin_rmpg_varpoord, 100 * lin_rmpg_convey[, 1])
```

```
## [1] TRUE
```

```
# variances do not match exactly
attr(svyrmpg( ~ eqincome , des_eusilc) , 'var') * 10000
```

```
##          eqincome
## eqincome 0.332234
```

```
varpoord_rmpg_calculation$all_result$var
```

```
## [1] 0.3316454
```

```
# standard errors do not match exactly
varpoord_rmpg_calculation$all_result$se
```

```
## [1] 0.5758866
```

```
SE(svyrmpg( ~ eqincome , des_eusilc)) * 100
```

```
##           eqincome
## eqincome 0.5763974
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svyrmpg` and `vardpoor::linrmpg` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
```

```
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )

# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)



# matches
stopifnot(all.equal(
  attr(svyrmpg( ~ eqincome , des_eusilc_ultimate_cluster) , 'var')[1] * 10000 ,
  varpoord_rmpg_calculation$all_result$var
))

# matches
stopifnot(all.equal(SE(
  svyrmpg(~ eqincome , des_eusilc_ultimate_cluster)
)[1] * 100 ,
varpoord_rmpg_calculation$all_result$se))
```

For additional usage examples of `svyrmpg`, type `?convey::svyrmpg` in the R console.

2.5 Median Income Below the At Risk of Poverty Threshold (svypoormed)

Median income below the at-risk-of-poverty-threshold (POORMED) is median of incomes of people having the income below the ARPT:

$$poormed = median\{y_i; y_i < arpt\}$$

The details of the linearization of the POORMED are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a POORMED coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the vardpoor library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the poormed coefficient
# using the R vardpoor library
varpoord_poormed_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
```

```
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # poormed coefficient function
    type = "linpoormed",

    # get linearized variable
    outp_lin = TRUE

  )



# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )

# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_poormed_calculation$all_result$value
```

```
## [1] 8803.735
```

```
coef(svypoormed( ~ eqincome , des_eusilc))
```

```
## eqincome
## 8803.735
```

```
# linearized variables do match
# vardpoor
lin_poormed_varpoord <-
  varpoord_poormed_calculation$lin_out$lin_poormed
# convey
lin_poormed_convey <-
  attr(svypoormed( ~ eqincome , des_eusilc), "lin")

# check equality
all.equal(lin_poormed_varpoord, lin_poormed_convey)
```

```
## [1] TRUE
```

```
# variances do not match exactly
attr(svypoormed( ~ eqincome , des_eusilc) , 'var')
```

```
##          eqincome
## eqincome  5311.47
```

```
varpoord_poormed_calculation$all_result$var
```

```
## [1] 5302.086
```

```
# standard errors do not match exactly
varpoord_poormed_calculation$all_result$se
```

```
## [1] 72.81542
```

```
SE(svypoormed( ~ eqincome , des_eusilc))
```

```
##          eqincome
## eqincome 72.87983
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svypoormed` and `vardpoor::linpoormed` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )

# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)


# matches
stopifnot(all.equal(
  attr(svypoormed(~ eqincome , des_eusilc_ultimate_cluster) , 'var')[1],
  varpoord_poormed_calculation$all_result$var
))

# matches
stopifnot(all.equal(
  SE(svypoormed(~ eqincome , des_eusilc_ultimate_cluster))[1],
  varpoord_poormed_calculation$all_result$se
))
```

For additional usage examples of `svypoormed`, type `?convey::svypoormed` in the R console.

2.6 Foster-Greer-Thorbecke class (svyfgt, svyfgtdec)

Foster et al. (1984) proposed a family of indicators to measure poverty. This class of *FGT* measures, can be defined as

$$p = \frac{1}{N} \sum_{k \in U} h(y_k, \theta),$$

where

$$h(y_k, \theta) = \left[\frac{(\theta - y_k)}{\theta} \right]^{\gamma} \delta \left\{ y_k \leq \theta \right\},$$

where: θ is the poverty threshold; δ the indicator function that assigns value 1 if the condition $\{y_k \leq \theta\}$ is satisfied and 0 otherwise, and γ is a non-negative constant.

If $\gamma = 0$, the FGT(0) equals the poverty headcount ratio, which accounts for the spread of poverty. If $\gamma = 1$, FGT(1) is the mean of the normalized income shortfall of the poor. By doing so, the measure takes into account both the spread and the intensity of poverty. When $\gamma = 2$, the relative weight of larger shortfalls increases even more, which yields a measure that accounts for poverty severity, i.e., the inequality among the poor. This way, a transfer from a poor person to an even poorer person would reduce the FGT(2).

Although Foster et al. (1984) already presented a decomposition for the FGT(2) index, Aristondo et al. (2010) provided a general formula that decomposes the FGT(γ) for any $\gamma \geqslant 2$. Put simply, any such FGT(γ) index can be seen as function of the headcount ratio, the average normalized income gap among the poor, and a generalized entropy index of the normalized income gaps among poor. In mathematical terms,

$$FGT_{\gamma} = FGT_0 \cdot I^{\gamma} \cdot [1 + (\gamma^2 - \gamma) GEI^*_{\gamma}], \ \gamma \geq 2$$

where I is the average normalized income gap among the poor and GEI^*_{γ} is a generalized entropy index of such income gaps among the poor.

This result is particularly useful, as one can attribute cross-sectional differences of a FGT index to differences in the spread, depth, and inequality of poverty.

The FGT poverty class and its decomposition is implemented in the `convey` library by the function `svyfgt` and `svyfgtdec`, respectively. The argument `thresh_type` of this function defines the type of poverty threshold adopted. There are three possible choices:

1. `abs` – fixed and given by the argument thresh_value.
2. `relq` – a proportion of a quantile fixed by the argument `proportion` and the quantile is defined by the argument `order`.
3. `relm` – a proportion of the mean fixed the argument `proportion`.

The quantile and the mean involved in the definition of the threshold are estimated for the whole population. When $\gamma = 0$ and $\theta = .6 * MED$, this measure is equal to the indicator ARPR computed by the function `svyarpr`. The linearization of the FGT(0) is presented in Berger and Skinner (2003).

Next, we give some examples of the function `svyfgt` to estimate the values of the FGT poverty index.

Consider first the poverty threshold fixed ($\gamma = 0$) in the value 10000. The headcount ratio (FGT0) is

```
svyfgt( ~ eqincome, des_eusilc, g = 0, abs_thresh = 10000)
```

```
            fgt0     SE
eqincome 0.11444 0.0027
```

The poverty gap ratio (FGT(1)) ($\gamma = 1$) index for the poverty threshold fixed at the same value is

```
svyfgt( ~ eqincome, des_eusilc, g = 1, abs_thresh = 10000)
```

```
             fgt1     SE
eqincome 0.032085 0.0011
```

To estimate the FGT(0) with the poverty threshold fixed at $0.6 * MED$, we first fix the argument `type_thresh="relq"` and then use the default values for `percent` and `order`:

```
svyfgt( ~ eqincome, des_eusilc, g = 0, type_thresh = "relq")
```

```
            fgt0     SE
eqincome 0.14444 0.0028
```

This matches the estimate obtained by:

```
svyarpr( ~ eqincome, design = des_eusilc, .5, .6)
```

```
            arpr     SE
eqincome 0.14444 0.0028
```

To estimate the poverty gap ratio with the poverty threshold equal to $0.6 * MEAN$, we use:

```
svyfgt( ~ eqincome, des_eusilc, g = 1, type_thresh = "relm")
```

```
             fgt1     SE
eqincome 0.051187 0.0011
```

A replication example

In July 2006, Jenkins (2008) presented at the North American Stata Users' Group Meetings on the stata Atkinson Index command. The example below reproduces those statistics.

In order to match the presentation's results using the `svyfgt` function from the `convey` library, the poverty threshold was considered absolute despite being directly estimated from the survey sample. This effectively treats the variance of the estimated poverty threshold as zero; `svyfgt` does not account for the uncertainty of the poverty threshold when the level has been stated as absolute with the `abs_thresh=` parameter. In general, we would instead recommend using either `relq` or `relm` in the `type_thresh=` parameter in order to account for the added uncertainty of the poverty threshold calculation. This example serves only to show that `svyfgt` behaves properly as compared to other software.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the foreign library
library(foreign)

# create a temporary file on the local disk
tf <- tempfile()

# store the location of the presentation file
presentation_zip <-
  "http://repec.org/nasug2006/nasug2006_jenkins.zip"

# download jenkins' presentation to the temporary file
download.file(presentation_zip , tf , mode = 'wb')
```

```
# unzip the contents of the archive
presentation_files <- unzip(tf , exdir = tempdir())

# load the institute for fiscal studies' 1981, 1985, and 1991 data.frame objects
x81 <-
  read.dta(grep("ifs81" , presentation_files , value = TRUE))
x85 <-
  read.dta(grep("ifs85" , presentation_files , value = TRUE))
x91 <-
  read.dta(grep("ifs91" , presentation_files , value = TRUE))

# NOTE: we recommend using ?convey::svyarpt rather than this unweighted calculation #

# calculate 60% of the unweighted median income in 1981
unwtd_arpt81 <- quantile(x81$eybhc0 , 0.5) * .6

# calculate 60% of the unweighted median income in 1985
unwtd_arpt85 <- quantile(x85$eybhc0 , 0.5) * .6

# calculate 60% of the unweighted median income in 1991
unwtd_arpt91 <- quantile(x91$eybhc0 , 0.5) * .6

# stack each of these three years of data into a single data.frame
x <- rbind(x81 , x85 , x91)
```

Replicate the author's survey design statement from stata code..

```
. ge poor = (year==1981)*(x < $z_81)+(year==1985)*(x < $z_85)+(year==1991)*(x < $z_91)
. * account for clustering within HHs
. svyset hrn [pweight = wgt]
```

.. into R code:

```
# initiate a linearized survey design object
y <- svydesign( ~ hrn , data = x , weights = ~ wgt)

# immediately run the `convey_prep` function on the survey design
z <- convey_prep(y)
```

Replicate the author's headcount ratio results with stata..

```
. svy: mean poor if year == 1981
(running mean on estimation sample)
```

```
Survey: Mean estimation

Number of strata =       1          Number of obs    =    9772
Number of PSUs   =    7476          Population size  = 5.5e+07
                                    Design df        =    7475

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |   .1410125   .0044859       .132219     .149806
--------------------------------------------------------------

. svy: mean poor if year == 1985
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       1          Number of obs    =    8991
Number of PSUs   =    6972          Population size  = 5.5e+07
                                    Design df        =    6971

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |    .137645   .0046531      .1285235    .1467665
--------------------------------------------------------------

. svy: mean poor if year == 1991
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       1          Number of obs    =    6468
Number of PSUs   =    5254          Population size  = 5.6e+07
                                    Design df        =    5253

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        poor |   .2021312   .0062077      .1899615    .2143009
--------------------------------------------------------------
```

..using R code:

```
headcount_81 <-
  svyfgt( ~ eybhc0 ,
          subset(z , year == 1981) ,
          g = 0 ,
          abs_thresh = unwtd_arpt81)

stopifnot(round(coef(headcount_81) , 5) == .14101)
stopifnot(round(SE(headcount_81) , 5) == .00449)

headcount_81_ci <-
  confint(headcount_81 , df = degf(subset(z , year == 1981)))

stopifnot(round(headcount_81_ci[1] , 5) == .13222)
stopifnot(round(headcount_81_ci[2] , 5) == .14981)

headcount_85 <-
  svyfgt(~ eybhc0 ,
         subset(z , year == 1985) ,
         g = 0 ,
         abs_thresh = unwtd_arpt85)


stopifnot(round(coef(headcount_85) , 5) == .13764)
stopifnot(round(SE(headcount_85) , 5) == .00465)

headcount_85_ci <-
  confint(headcount_85 , df = degf(subset(z , year == 1985)))

stopifnot(round(headcount_85_ci[1] , 5) == .12852)
stopifnot(round(headcount_85_ci[2] , 5) == .14677)

headcount_91 <-
  svyfgt(~ eybhc0 ,
         subset(z , year == 1991) ,
         g = 0 ,
         abs_thresh = unwtd_arpt91)


stopifnot(round(coef(headcount_91) , 5) == .20213)
stopifnot(round(SE(headcount_91) , 5) == .00621)

headcount_91_ci <-
  confint(headcount_91 , df = degf(subset(z , year == 1991)))
```

```
stopifnot(round(headcount_91_ci[1] , 5) == .18996)
stopifnot(round(headcount_91_ci[2] , 5) == .21430)
```

Confirm this replication applies for the normalized poverty gap as well, comparing stata code..

```
. ge ngap = poor*($z_81- x)/$z_81 if year == 1981

. svy: mean ngap if year == 1981
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =       1          Number of obs    =    9772
Number of PSUs   =    7476          Population size  = 5.5e+07
                                    Design df        =    7475

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
        ngap |   .0271577   .0013502      .0245109    .0298044
--------------------------------------------------------------
```

..to R code:

```
norm_pov_81 <-
  svyfgt( ~ eybhc0 ,
          subset(z , year == 1981) ,
          g = 1 ,
          abs_thresh = unwtd_arpt81)


stopifnot(round(coef(norm_pov_81) , 5) == .02716)
stopifnot(round(SE(norm_pov_81) , 5) == .00135)

norm_pov_81_ci <-
  confint(norm_pov_81 , df = degf(subset(z , year == 1981)))

stopifnot(round(norm_pov_81_ci[1] , 5) == .02451)
stopifnot(round(norm_pov_81_ci[2] , 5) == .02980)
```

To provide a example for our code, we proceed with a Monte Carlo experiment. Using the `eusilcP` data from the `simPop` package (Templ et al., 2017), we can compute the actual values of the FGT components for that population:

```
# load libraries
library(sampling)
library(survey)
library(convey)
library(parallel)

# load pseudo population data
data("eusilcP" , package = "simPop")

# compute population value of the FGT(2) index decomposition
inc.all <- eusilcP$eqIncome
gfun <- function(y , thresh)
  (thresh - y) / thresh
hfun <-
  function(y , thresh , g)
    (((thresh - y) / thresh) ^ g) * (y <= thresh)
fgt2 <- mean(hfun(inc.all , 10000 , 2) , na.rm = TRUE)
fgt1 <- mean(hfun(inc.all , 10000 , 1) , na.rm = TRUE)
fgt0 <- mean(hfun(inc.all , 10000 , 0) , na.rm = TRUE)
igr.fgt <-
  mean(hfun(inc.all , 10000 , 1)[inc.all <= 10000] , na.rm = TRUE)
gei.poor <-
  convey:::CalcGEI(gfun(inc.all , 10000) , ifelse(inc.all < 10000 , 1 , 0) , 2)
theta.pop <-
  c(
    "fgt2" = fgt2 ,
    "fgt0" =  fgt0 ,
    "fgt1" =  fgt1 ,
    "igr" = igr.fgt ,
    "gei(poor;epsilon=2)" =  gei.poor
  )
theta.pop
```

```
##                fgt2                fgt0                fgt1                 igr
##          0.01571192          0.11412691          0.03259544          0.28560699
## gei(poor;epsilon=2)
##          0.34386588
```

Now, to study the distribution of the estimator under a particular sampling design, we select 5000 samples under one-stage cluster sampling of 100 households using the `cluster` function from the `sampling` package (Tillé & Matei, 2021), and use the `svyfgtdec` function to estimate the FGT components for each of those samples:

```r
# define the number of monte carlo replicates
mc.rep <- 5000L



# simulation function
fgtdec_sim_fun <- function(this.iter) {
  set.seed(this.iter)

  library(sampling)
  library(survey)
  library(convey)

  # load pseudo population data
  data("eusilcP" , package = "simPop")


  # compute size-like variable for PPS sampling design
  eusilcP$aux <-
    log(ifelse(eusilcP$eqIncome >= 1000 , eusilcP$eqIncome , 1000))


  # select sample
  tt <-
    sampling::cluster(
      data = eusilcP[sample.int(nrow(eusilcP) , nrow(eusilcP) , replace = FALSE) , ] ,
      clustername = "hid" ,
      size = 1000L ,
      method = "systematic" ,
      pik = eusilcP$aux
    )

  # collect data
  this.sample <- getdata(eusilcP , tt)

  # create survey design object
  this.desobj <-
    svydesign(
      ids = ~ hid ,
      probs = ~ Prob ,
      data = this.sample ,
      nest = FALSE
    )

  # prepare for convey functions
```

```
  this.desobj <- convey_prep(this.desobj)

  # compute estimates
  svyfgtdec( ~ eqIncome , this.desobj , abs_thresh = 10000 , g = 2)

}

# run replications
cl <- makeCluster(detectCores() - 1)

fgtdec.estimate.list <-
  clusterApply(cl, seq_len(mc.rep) , fgtdec_sim_fun)

stopCluster(cl)
```

The PRB of each component is estimated using the code below. Notice that PRBs are relatively small, with absolute values below 1%, with the largest bias in the GEI index component.

```
# repeat true population values
(theta.pop)
```

```
##                fgt2                fgt0                fgt1                 igr
##          0.01571192          0.11412691          0.03259544          0.28560699
## gei(poor;epsilon=2)
##          0.34386588
```

```
# estimate the expected values of the components estimators
# using the average of the estimates
(theta.exp <- rowMeans(sapply(fgtdec.estimate.list , coef)))
```

```
##                fgt2                fgt0                fgt1                 igr
##          0.01577844          0.11432776          0.03269125          0.28591312
## gei(poor;epsilon=2)
##          0.34234370
```

```
# estimate the percentage relative bias
(percentage_relative_bias <- 100 * (theta.exp / theta.pop - 1))
```

```
##                fgt2                fgt0                fgt1                 igr
##           0.4233956           0.1759893           0.2939132           0.1071867
## gei(poor;epsilon=2)
##          -0.4426679
```

```
stopifnot(abs(percentage_relative_bias) < 1)
```

For the variance estimators, we estimate the PRB using the code below. Note that the bias is still relatively small, with absolute values of the PRB below 5%.

```
# estimate the variance of the components estimators
# using the empirical variance of the estimates
(vartheta.popest <-
   diag(var(t(
     sapply(fgtdec.estimate.list , coef)
   ))))
```

```
##                fgt2                fgt0                fgt1                 igr
##        6.447720e-06        1.015546e-04        1.468049e-05        4.753320e-04
## gei(poor;epsilon=2)
##        1.841360e-03
```

```
# estimate the expected value of the Watts index variance estimator
# using the (estimated) expected value of the variance estimates
(vartheta.exp <-
    rowMeans(sapply(fgtdec.estimate.list , function(z)
      diag(vcov(
        z
      )))))
```

```
## [1] 6.462904e-06 1.014859e-04 1.473047e-05 4.923412e-04 1.911158e-03
```

```
# estimate the percentage relative bias of the variance estimators
(percentage_relative_bias_variance <- 100 *  (vartheta.exp / vartheta.popest - 1))
```

```
##                fgt2                fgt0                fgt1                 igr
##          0.23548684         -0.06759586          0.34047392          3.57838870
## gei(poor;epsilon=2)
##          3.79052600
```

```
stopifnot(abs(percentage_relative_bias_variance) < 5)
```

Regarding the MSE, the squared bias accounts for less than 1% of the MSE. This means that, with a good estimate of the variance, we should be able to have a good approximation for the MSE.

```
# estimate MSE
theta.bias2 <- (theta.exp - theta.pop) ^ 2
(theta.mse <- theta.bias2 + vartheta.popest)
```

```
##                fgt2                fgt0                fgt1                 igr
##        6.472967e-06        1.035216e-04        1.489719e-05        4.941012e-04
## gei(poor;epsilon=2)
##        1.842408e-03
```

```
# squared bias component of the MSE
( squared_bias_over_mse <- 100 * (theta.bias2 / theta.mse) )
```

```
##                fgt2                fgt0                fgt1                 igr
##        0.0024344376        0.0021033034        0.0002215282        0.0066399605
## gei(poor;epsilon=2)
##        0.1389369917
```

```
stopifnot( squared_bias_over_mse < 1 )
```

The CIs based on the normal approximation work reasonably well for all components. The code below shows that the coverage rates are close to the 95% nominal coverage rate.

```
# estimate confidence intervals of the Watts index
# for each of the samples
est.coverage <-
  sapply(fgtdec.estimate.list, function(this.stat)
    confint(this.stat)[, 1] <= theta.pop &
      confint(this.stat)[, 2] >= theta.pop)

# evaluate empirical coverage
stopifnot( abs( rowMeans(est.coverage) - 0.95 ) < 0.015 )
```

For additional usage examples of `svyfgt` and `svyfgtdec`, type `?convey::svyfgt` or `?convey::svyfgtdec` in the R console.

2.7 Watts poverty measure (svywatts, svywattsdec)

The measure proposed in Watts (1968) satisfies a number of desirable poverty measurement axioms and is known to be one of the first distribution-sensitive poverty measures, as noted by Haughton and Khandker (2009). It is defined as:

$$Watts = \frac{1}{N}\sum_{i\in U}\log\left(\frac{y_i}{\theta}\right)\delta(y_i \leqslant \theta).$$

Morduch (1998) points out that the Watts poverty index can provide an estimate of the expected time to exit poverty. Given the expected growth rate of income per capita among the poor, g, the expected time taken to exit poverty T_θ would be:

$$T_\theta = \frac{Watts}{g}.$$

The Watts poverty index also has interesting decomposition properties. Blackburn (1989) proposed a decomposition for the Watts poverty index, rewriting it in terms of the headcount ratio, the Watts poverty gap ratio and the mean log deviaton of poor incomes[1]. Mathematically,

$$Watts = FGT_0(I_w + L_*)$$

where $I_w = \log(\theta/\mu_*)$ is the Watts poverty gap ratio[2] and L_* is the mean log deviation of incomes among the poor. This can be estimated using the `svywattsdec` function.

This result can also be interpreted as a decomposition of the time taken to exit poverty, since

$$\begin{aligned} T_\theta &= \frac{Watts}{g} \\ &= \frac{FGT_0}{g}(I_w + L_*) \end{aligned}$$

As Morduch (1998) points out, if the income among the poor is equally distributed (i.e., $L_* = 0$), the time taken to exit poverty is simply $FGT_0 I_w/g$. Therefore, $FGT_0 L_*/g$ can be seen as the additional time needed to exit poverty as a result of the inequality among the poor.

A replication example

To provide a example for our code, we proceed with a Monte Carlo experiment. Using the `eusilcP` data from the `simPop` package (Templ et al., 2017), we can compute the actual value of the Watts index for that population:

[1] The mean log deviation (also known as Theil-L or Bourguignon-Theil index) is an inequality measure of the generalized entropy class. The family of generalized entropy indices is discussed in the next chapter.

[2] μ_* stands for the average income among the poor.

```r
# load libraries
library(sampling)
library(survey)
library(convey)
library(parallel)

# load pseudo population data
data("eusilcP" , package = "simPop")

# compute population value of the Watts index decomposition
inc.pos <- eusilcP$eqIncome[eusilcP$eqIncome > 0]
(theta.pop <-
    mean(ifelse(inc.pos <= 10000 , log(10000 /  inc.pos) , 0) , na.rm = TRUE))
```

```
## [1] 0.05025374
```

Now, to study the distribution of the estimator under a particular sampling design, we select 5000 samples under one-stage cluster sampling of 100 households using the `cluster` function from the `sampling` package (Tillé & Matei, 2021), and use the `svywatts` function to estimate the Watts index for each of those samples:

```r
# define the number of monte carlo replicates
mc.rep <- 5000L


# simulation function
watts_sim_fun <- function(this.iter) {
  set.seed(this.iter)

  library(survey)
  library(convey)
  library(sampling)


  # load pseudo population data
  data("eusilcP" , package = "simPop")

  # compute size-like variable for PPS sampling design
  eusilcP$aux <-
    log(ifelse(eusilcP$eqIncome >= 1000 , eusilcP$eqIncome , 1000))

  # select sample
  tt <-
```

```
    sampling::cluster(
      data = eusilcP[sample.int(nrow(eusilcP) , nrow(eusilcP) , replace = FALSE) ,] ,
      clustername = "hid" ,
      size = 1000L ,
      method = "systematic" ,
      pik = eusilcP$aux
    )

  # collect data
  this.sample <- getdata(eusilcP , tt)

  # create survey design object
  this.desobj <-
    svydesign(
      ids = ~ hid ,
      probs = ~ Prob ,
      data = this.sample ,
      nest = FALSE
    )

  # prepare for convey functions
  this.desobj <- convey_prep(this.desobj)

  # filter positive incomes
  this.desobj <- subset(this.desobj , eqIncome > 0)

  # compute estimates
  svywatts( ~ eqIncome , this.desobj , abs_thresh = 10000)

}

# run replications
cl <- makeCluster(detectCores() - 1)

watts.estimate.list <-
  clusterApply(cl, seq_len(mc.rep) , watts_sim_fun)

stopCluster(cl)
```

Then, we evaluate the Percentage Relative Bias (PRB) of the Watts index estimator. Under this scenario, the PRB of the Watts index estimator is 0.3772%.

```
# estimate the expected value of the Watts index estimator
# using the average of the estimates
(theta.exp <- mean(sapply(watts.estimate.list , coef)))
```

```
## [1] 0.05044329
```

```
# estimate the percentage relative bias
(percentage_relative_bias <- 100 * (theta.exp / theta.pop - 1) )
```

```
## [1] 0.377195
```

```
stopifnot( round( percentage_relative_bias , 4 ) == 0.3772)
```

For the variance estimator, we have:

```
# estimate the variance of the Watts index estimator
# using the empirical variance of the estimates
(vartheta.popest <- var(sapply(watts.estimate.list , coef)))
```

```
## [1] 6.141434e-05
```

```
# estimate the expected value of the Watts index variance estimator
# using the (estimated) expected value of the variance estimates
(vartheta.exp <- mean(sapply(watts.estimate.list , vcov)))
```

```
## [1] 6.100902e-05
```

```
# estimate the percentage relative bias of the variance estimator
(percentage_relative_bias_variance <- 100 *  (vartheta.exp / vartheta.popest - 1))
```

```
## [1] -0.6599717
```

```
stopifnot(round(percentage_relative_bias_variance, 4) == -0.6600)
```

Under this scenario, the PRB of the Watts index variance estimator is -0.6600%.

Our simulations show that the Squared Bias of this estimator accounts for less than 0.1% of its Mean Squared Error:

```
theta.bias2 <- (theta.exp - theta.pop) ^ 2
theta.mse <- theta.bias2 + vartheta.popest
(squared_bias_over_mse <- 100 * (theta.bias2 / theta.mse))
```

```
## [1] 0.05847158
```

```
stopifnot(squared_bias_over_mse < 0.1)
```

Next, we evaluate the Percentage Coverage Rate (PCR). In theory, under repeated sampling, the estimated 95% CIs should cover the population parameter approximately 95% of the time. We can evaluate that using:

```
# estimate confidence intervals of the Watts index
# for each of the samples
est.coverage <-
  sapply(watts.estimate.list, function(this.stat)
    confint(this.stat)[, 1] <= theta.pop &
      confint(this.stat)[, 2] >= theta.pop)

# evaluate empirical coverage
(empirical_coverage <- mean(est.coverage))
```

```
## [1] 0.9268
```

```
stopifnot(abs(empirical_coverage  - 0.95) < 0.025)
```

Our coverages are not too far from the nominal coverage level of 95%.

For the Watts index decomposition, we start by computing the (true) population values of the components:

```
# compute population value of the Watts index decomposition
inc.pos <- eusilcP$eqIncome[eusilcP$eqIncome > 0]
wdec1 <-
  mean(ifelse(inc.pos <= 10000 , log(10000 /  inc.pos) , 0) , na.rm = TRUE)
wdec2 <- mean(inc.pos <= 10000 , na.rm = TRUE)
mu.poor <- mean(inc.pos [inc.pos <= 10000])
wdec3 <- log(10000 / mu.poor)
wdec4 <-
  -mean(log(inc.pos[inc.pos <= 10000] / mu.poor) , na.rm = TRUE)
theta.pop <-
  c(
    "watts" = wdec1 ,
    "fgt0" =  wdec2 ,
    "watts pov. gap ratio" = wdec3 ,
    "theil(poor)" =  wdec4
  )
theta.pop
```

```
##                watts                 fgt0 watts pov. gap ratio
```

```
##           0.05025374           0.11399096           0.33497664
##          theil(poor)
##           0.10588056
```

Then, using the same sampling strategy of the `svywatts`, we compute the `svywattsdec` for each sample:

```
# simulation function
wattsdec_sim_fun <- function(this.iter) {

  set.seed(this.iter)

  library(survey)
  library(convey)
  library(sampling)


  # load pseudo population data
  data("eusilcP" , package = "simPop")

  # compute size-like variable for PPS sampling design
  eusilcP$aux <-
    log(ifelse(eusilcP$eqIncome >= 1000 , eusilcP$eqIncome , 1000))

  # select sample
  tt <-
    sampling::cluster(
      data = eusilcP[sample.int(nrow(eusilcP) , nrow(eusilcP) , replace = FALSE) , ] ,
      clustername = "hid" ,
      size = 1000L ,
      method = "systematic" ,
      pik = eusilcP$aux
    )

  # collect data
  this.sample <- getdata(eusilcP , tt)

  # create survey design object
  this.desobj <-
    svydesign(
      ids = ~ hid ,
      probs = ~ Prob ,
      data = this.sample ,
      nest = FALSE
    )
```

```
  # prepare for convey functions
  this.desobj <- convey_prep(this.desobj)

  # filter positive incomes
  this.desobj <- subset(this.desobj , eqIncome > 0)

  # compute estimates
  svywattsdec(~ eqIncome , this.desobj , abs_thresh = 10000)

}

# run replications
cl <- makeCluster(detectCores() - 1)

wattsdec.estimate.list <-
  clusterApply(cl, seq_len(mc.rep) , wattsdec_sim_fun)

stopCluster(cl)
```

The PRB of each component is estimated using the code below. Notice that PRBs are relatively small, with absolute values below 1%, with the largest bias in the Theil index component.

```
# repeat true population values
(theta.pop)
```

```
##                watts                 fgt0 watts pov. gap ratio
##           0.05025374           0.11399096           0.33497664
##          theil(poor)
##           0.10588056
```

```
# estimate the expected values of the components estimators
# using the average of the estimates
(theta.exp <- rowMeans(sapply(wattsdec.estimate.list , coef)))
```

```
##                watts                 fgt0 watts pov. gap ratio
##           0.05044329           0.11418992           0.33584759
##          theil(poor)
##           0.10584416
```

```
# estimate the percentage relative bias
(percentage_relative_bias <- 100 * (theta.exp / theta.pop - 1))
```

```
##               watts                fgt0 watts pov. gap ratio
##           0.37719501           0.17453750           0.26000223
##          theil(poor)
##          -0.03437576
```

```
stopifnot(abs(percentage_relative_bias) < 1)
```

For the variance estimators, we estimate the PRB using the code below. Note that the bias is still relatively small, with absolute values of the PRB below 1%.

```
# estimate the variance of the components estimators
# using the empirical variance of the estimates
(vartheta.popest <-
   diag(var(t(
     sapply(wattsdec.estimate.list , coef)
   ))))
```

```
##               watts                fgt0 watts pov. gap ratio
##         6.141434e-05         1.015260e-04         9.241259e-04
##          theil(poor)
##         1.108525e-03
```

```
# estimate the expected value of the Watts index variance estimator
# using the (estimated) expected value of the variance estimates
(vartheta.exp <-
    rowMeans(sapply(wattsdec.estimate.list , function(z)
      diag(vcov(
        z
      )))))
```

```
##               watts                fgt0 watts pov. gap ratio
##         6.100902e-05         1.013968e-04         9.613831e-04
##          theil(poor)
##         1.070018e-03
```

```
# estimate the percentage relative bias of the variance estimators
(percentage_relative_bias <-
    100 *  (vartheta.exp / vartheta.popest - 1))
```

```
##               watts                fgt0 watts pov. gap ratio
##           -0.6599717           -0.1273107            4.0316155
##          theil(poor)
##           -3.4736907
```

```
stopifnot(abs(percentage_relative_bias[1]) < 1)
```

Regarding the MSE, the squared bias accounts for less than 0.1% of the MSE. This means that, with a good estimate of the variance, we should be able to have a good approximation for the MSE.

```
# estimate MSE
theta.bias2 <- (theta.exp - theta.pop) ^ 2
(theta.mse <- theta.bias2 + vartheta.popest)
```

```
##                watts                 fgt0 watts pov. gap ratio
##         6.145027e-05         1.015656e-04         9.248844e-04
##          theil(poor)
##         1.108526e-03
```

```
# squared bias component of the MSE
(squared_bias_over_mse <- 100 * (theta.bias2 / theta.mse))
```

```
##                watts                 fgt0 watts pov. gap ratio
##         0.0584715789         0.0389737000         0.0820154610
##          theil(poor)
##         0.0001195063
```

```
stopifnot(squared_bias_over_mse < 0.1)
```

However, the CIs based on the normal approximation might not work very well for some components. The code below shows that coverage rate for the Theil index component differs significantly from the 95% nominal coverage rate.

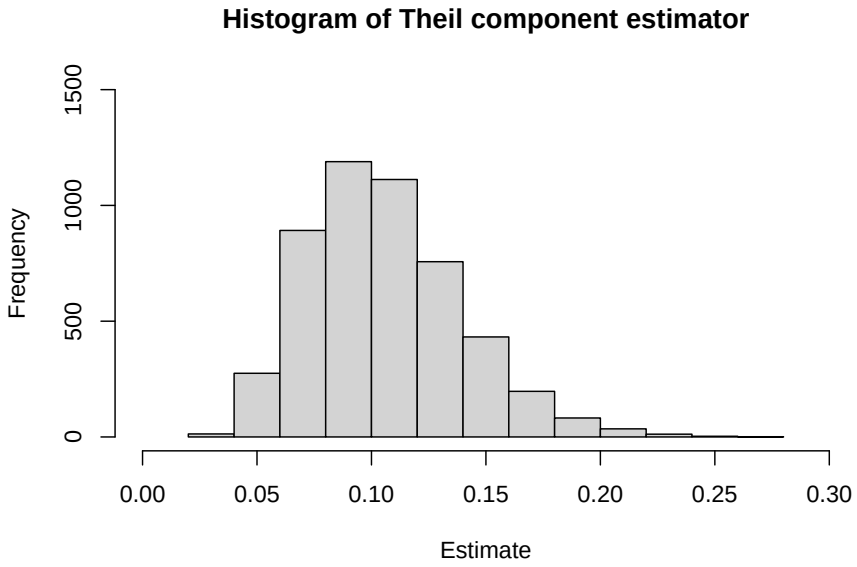
```
# estimate confidence intervals of the Watts index
# for each of the samples
est.coverage <-
  sapply(wattsdec.estimate.list, function(this.stat)
    confint(this.stat)[, 1] <= theta.pop &
      confint(this.stat)[, 2] >= theta.pop)

# evaluate empirical coverage
(empirical_coverage <- rowMeans(est.coverage))
```

```
##                watts                 fgt0 watts pov. gap ratio
##               0.9268               0.9494               0.9476
##          theil(poor)
##               0.8452
```

```
stopifnot(abs(empirical_coverage - 0.95) < 0.15)
```

One of the reasons for this is that the sample might not be large enough for the CLT to hold. The distribution of the estimator shows substantial asymmetry, which would be a problem for the normal approximation.

```
hist(
  sapply(wattsdec.estimate.list , coef)[4, ] ,
  main = "Histogram of Theil component estimator" ,
  xlim = c(0, .30) ,
  ylim = c(0 , 1500) ,
  xlab = "Estimate"
)
```



For additional usage examples of `svywatts` and `svywattsdec`, type `?convey::svywatts` or `?convey::svywattsdec` in the R console.

Chapter 3

Inequality Measurement

Another problem faced by societies is inequality. Economic inequality can have several different meanings, including (but not limited to) income, education, resources, opportunities, and well-being. Usually, studies on economic inequality focus on income distribution.

Most inequality data comes from censuses and household surveys. Therefore, in order to produce reliable estimates from survey data, procedures that account for complex sampling designs are necessary.

This chapter presents the inequality measures available in the `convey` library with replication examples where possible. It starts with an attempt to measure inequality between two groups of a population, then summarizes the general idea of inequality indices by discussing techniques like the quintile share ratio, the Lorenz curve, and the commonly-used Gini coefficient. Next, this chapter discusses the concept of entropy and presents inequality measures (and their decompositions) based on that method. This chapter concludes with a discussion regarding the tradeoffs of selecting among the various inequality measures.

3.1 The Gender Pay Gap (svygpg)

Although the Gender Pay Gap (GPG) is not an inequality measure in the usual sense, it can still be a useful instrument to evaluate the effects of gender discrimination. Put simply, it expresses the relative difference between the average hourly earnings of men and women, presenting this difference as a percentage of the average of hourly earnings of men. Like some other functions described in this text, the GPG can also be calculated using wealth or assets in place of earnings.

In mathematical terms, this index can be described as,

$$GPG = \frac{\bar{y}_{male} - \bar{y}_{female}}{\bar{y}_{male}}$$

As we can see from the formula, if there is no difference among classes, $GPG = 0$. Else, if $GPG > 0$, it means that the average hourly income received by women are GPG percent smaller than men's. For negative GPG, it means that women's hourly earnings are GPG percent larger than men's. In other words, the larger the GPG, larger is the shortfall of women's hourly earnings.

We can also develop a more straightforward idea: for every \$1 raise in men's hourly earnings, women's hourly earnings are expected to increase $\$(1 - GPG)$. For instance, assuming $GPG = 0.8$, for every \$1.00 increase in men's average hourly earnings, women's hourly earnings would increase only \$0.20.

The details of the linearization of the GPG are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a GPG coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the laeken library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# coerce the gender variable to numeric 1 or 2
```

```r
eusilc$one_two <- as.numeric(eusilc$rb090 == "female") + 1

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the gpg coefficient
# using the R vardpoor library
varpoord_gpg_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # gpg coefficient function
    type = "lingpg" ,

    # gender variable
    gender = "one_two",

    # get linearized variable
    outp_lin = TRUE
  )



# construct a survey.design
```

```
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )

# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_gpg_calculation$all_result$value
```

```
## [1] 7.645389
```

```
coef(svygpg( ~ eqincome , des_eusilc , sex = ~ rb090)) * 100
```

```
##  eqincome
## -8.278297
```

```
# linearized variables do match
# vardpoor
lin_gpg_varpoord <- varpoord_gpg_calculation$lin_out$lin_gpg
# convey
lin_gpg_convey <-
  attr(svygpg( ~ eqincome , des_eusilc, sex = ~ rb090), "lin")

# check equality
all.equal(lin_gpg_varpoord, 100 * lin_gpg_convey[, 1])
```

```
## [1] "Mean relative difference: 2.172419"
```

```
# variances do not match exactly
attr(svygpg( ~ eqincome , des_eusilc , sex = ~ rb090) , 'var') * 10000
```

```
##           eqincome
## eqincome 0.8926311
```

```
varpoord_gpg_calculation$all_result$var
```

```
## [1] 0.6482346
```

```
# standard errors do not match exactly
varpoord_gpg_calculation$all_result$se
```

```
## [1] 0.8051301
```

```
SE(svygpg( ~ eqincome , des_eusilc , sex = ~ rb090)) * 100
```

```
##           eqincome
## eqincome 0.9447916
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svygpg` and `vardpoor::lingpg` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )
```

```
# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)

# matches
attr(svygpg( ~ eqincome , des_eusilc_ultimate_cluster , sex = ~ rb090) ,
     'var') * 10000
```

```
##           eqincome
## eqincome 0.8910413
```

```
varpoord_gpg_calculation$all_result$var
```

```
## [1] 0.6482346
```

```
# matches
varpoord_gpg_calculation$all_result$se
```

```
## [1] 0.8051301
```

```
SE(svygpg( ~ eqincome , des_eusilc_ultimate_cluster , sex = ~ rb090)) * 100
```

```
##           eqincome
## eqincome 0.9439499
```

For additional usage examples of `svygpg`, type `?convey::svygpg` in the R console.

3.2 Quintile Share Ratio (svyqsr)

Unlike the previous measure, the Quintile Share Ratio (QSR) is an inequality measure in itself, depending only on the income distribution to evaluate the degree of inequality. By definition, it can be described as the ratio between the income share held by the richest 20% and the poorest 20% of the population.

In plain terms, it expresses how many times the wealthier part of the population are richer than the poorest part. For instance, a $QSR = 4$ implies that the upper class takes home 4 times as much of the total income as the poor.

The QSR can be modified to a more general function of percentile share ratios. For instance, Cobham et al. (2015) argues for using the Palma index, defined as the ratio between the share of the 10% richest over the share held by the

poorest 40%. There are actually two ways to compute the Palma ratio with the `convey` package. One is using `convey::svyqsr` and the other is using `convey::svylorenz` - they won't match exactly because of different estimators, but are asymptotically equal. Since the Palma ratio is the top 10% divided by the bottom 40%, in `convey::svyqsr` this could be achieved using `alpha1 = .40` and `alpha2 = .90`. Note that the `vardpoor::linqsr` function only accepts a single `alpha` parameter (defaulting to `0.2` and `1 - 0.2`), meaning the Palma index cannot presently be computed using that function.

The details of the linearization of the QSR are discussed by Deville (1999) and Osier (2009).

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a QSR coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the laeken library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)

# calculate the qsr coefficient
# using the R vardpoor library
varpoord_qsr_calculation <-
  varpoord(
```

```
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # qsr coefficient function
    type = "linqsr",

    # poverty threshold range
    alpha = 20 ,

    # get linearized variable
    outp_lin = TRUE

  )



# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
  )
```

```
# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_qsr_calculation$all_result$value
```

```
## [1] 3.970004
```

```
coef(svyqsr( ~ eqincome , des_eusilc))
```

```
## eqincome
## 3.970004
```

```
# linearized variables do match
# vardpoor
lin_qsr_varpoord <- varpoord_qsr_calculation$lin_out$lin_qsr
# convey
lin_qsr_convey <-
  as.numeric(attr(svyqsr( ~ eqincome ,
                          des_eusilc ,
                          linearized = TRUE) ,
                  "linearized"))

# check equality
all.equal(lin_qsr_varpoord, lin_qsr_convey)
```

```
## [1] TRUE
```

```
# variances do not match exactly
attr(svyqsr( ~ eqincome , des_eusilc) , 'var')
```

```
##             eqincome
## eqincome 0.001810537
```

```
varpoord_qsr_calculation$all_result$var
```

```
## [1] 0.001807323
```

```
# standard errors do not match exactly
varpoord_qsr_calculation$all_result$se
```

```
## [1] 0.04251263
```

```
SE(svyqsr( ~ eqincome , des_eusilc))
```

```
##            eqincome
## eqincome 0.04255041
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svyqsr` and `vardpoor::linqsr` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )

# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)

# matches
attr(svyqsr( ~ eqincome , des_eusilc_ultimate_cluster) , 'var')
```

```
##             eqincome
## eqincome 0.001807323
```

```
varpoord_qsr_calculation$all_result$var
```

```
## [1] 0.001807323
```

```
# matches
varpoord_qsr_calculation$all_result$se
```

```
## [1] 0.04251263
```

```
SE(svyqsr( ~ eqincome , des_eusilc_ultimate_cluster))
```

```
##            eqincome
## eqincome 0.04251263
```

For additional usage examples of `svyqsr`, type `?convey::svyqsr` in the R console.

3.3 Lorenz Curve (svylorenz)

Though not an inequality measure in itself, the Lorenz curve is a classic instrument of distribution analysis. This function associates a cumulative share of the population to the share of the total income earned. In mathematical terms,

$$L(p) = \frac{\int_{-\infty}^{Q_p} yf(y)dy}{\int_{-\infty}^{+\infty} yf(y)dy}$$

where Q_p is the quantile p of the population.

The two extreme distributive cases are

- Perfect equality:
  - Every individual has the same income;
  - Every share of the population has the same share of the income;
  - Therefore, the reference curve is

$$L(p) = p \ \forall p \in [0, 1].$$

- Perfect inequality:
  - One individual concentrates all of society's income, while the other individuals have zero income;

– Therefore, the reference curve is

$$L(p) = \begin{cases} 0, & \forall p < 1 \\ 1, & \text{if } p = 1. \end{cases}$$

In order to evaluate the degree of inequality in a society, the analyst looks at the distance between the real curve and those two reference curves.

The estimator of this function was derived by Kovacevic and Binder (1997):

$$L(p) = \frac{\sum_{i \in S} w_i \cdot y_i \cdot \delta\{y_i \leq \widehat{Q}_p\}}{\widehat{Y}}, \ 0 \leq p \leq 1.$$

Yet, this formula is used to calculate specific points of the curve and their respective SEs. The formula to plot an approximation of the continuous empirical curve comes from Lerman and Yitzhaki (1989).

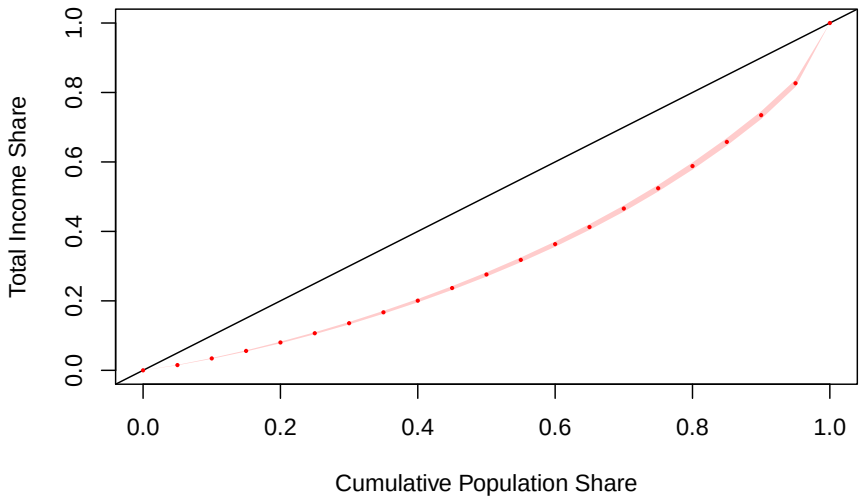
A replication example

In October 2016, (Jann, 2016) released a pre-publication working paper to estimate Lorenz and concentration curves using stata. The example below reproduces the statistics presented in his section 4.1.

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the stata-style webuse library
library(webuse)

# load the NLSW 1988 data
webuse("nlsw88")

# coerce that `tbl_df` to a standard R `data.frame`
nlsw88 <- data.frame(nlsw88)

# initiate a linearized survey design object
des_nlsw88 <- svydesign(ids = ~ 1 , data = nlsw88)
```

```
## Warning in svydesign.default(ids = ~1, data = nlsw88): No weights or
## probabilities supplied, assuming equal probability
```

```
# immediately run the `convey_prep` function on the survey design
des_nlsw88 <- convey_prep(des_nlsw88)

# estimates lorenz curve
result.lin <-
  svylorenz( ~ wage,
             des_nlsw88,
             quantiles = seq(0, 1, .05),
             na.rm = TRUE)
```




```
# note: most survey commands in R use Inf degrees of freedom by default
# stata generally uses the degrees of freedom of the survey design.
# therefore, while the degf() parameters passed to qt()
# serve to prove a precise replication of stata,
# it is generally not necessary
section_four_one <-
  data.frame(
    estimate = coef(result.lin) ,
    standard_error = SE(result.lin) ,
    ci_lower_bound =
      coef(result.lin) +
      SE(result.lin) *
      qt(0.025 , degf(subset(
        des_nlsw88 , !is.na(wage)
```

```
    ))) ,
  ci_upper_bound =
    coef(result.lin) +
    SE(result.lin) *
    qt(0.975 , degf(subset(
      des_nlsw88 , !is.na(wage)
    )))
)
```

	estimate	standard_error	ci_lower_bound	ci_upper_bound
L(0)	0.0000000	0.0000000	0.0000000	0.0000000
L(0.05)	0.0151060	0.0004159	0.0142904	0.0159216
L(0.1)	0.0342651	0.0007021	0.0328882	0.0356420
L(0.15)	0.0558635	0.0010096	0.0538836	0.0578434
L(0.2)	0.0801846	0.0014032	0.0774329	0.0829363
L(0.25)	0.1067687	0.0017315	0.1033732	0.1101642
L(0.3)	0.1356307	0.0021301	0.1314535	0.1398078
L(0.35)	0.1670287	0.0025182	0.1620903	0.1719670
L(0.4)	0.2005501	0.0029161	0.1948315	0.2062687
L(0.45)	0.2369209	0.0033267	0.2303971	0.2434447
L(0.5)	0.2759734	0.0037423	0.2686347	0.2833121
L(0.55)	0.3180215	0.0041626	0.3098585	0.3261844
L(0.6)	0.3633071	0.0045833	0.3543192	0.3722950
L(0.65)	0.4125183	0.0050056	0.4027021	0.4223345
L(0.7)	0.4657641	0.0054137	0.4551478	0.4763804
L(0.75)	0.5241784	0.0058003	0.5128039	0.5355529
L(0.8)	0.5880894	0.0062464	0.5758401	0.6003388
L(0.85)	0.6577051	0.0066148	0.6447333	0.6706769
L(0.9)	0.7346412	0.0068289	0.7212497	0.7480328
L(0.95)	0.8265786	0.0062686	0.8142857	0.8388715
L(1)	1.0000000	0.0000000	1.0000000	1.0000000

For additional usage examples of `svylorenz`, type `?convey::svylorenz` in the R console.

3.4 Gini index (svygini)

The Gini index (or Gini coefficient) is an attempt to express the inequality presented in the Lorenz curve as a single number. In essence, it is twice the area between the equality curve and the real Lorenz curve. Put simply:

$$G = 2\left(\int_0^1 pdp - \int_0^1 L(p)dp \right)$$
$$\therefore G = 1 - 2\int_0^1 L(p)dp$$

where $G = 0$ in case of perfect equality and $G = 1$ in the case of perfect inequality.

The estimator proposed by Osier (2009) is defined as:

$$\widehat{G} = \frac{2\sum_{i \in S} w_i r_i y_i - \sum_{i \in S} w_i y_i}{\hat{Y}}$$

The linearized formula of $\widehat{G}$ is used to calculate the SE.

A replication example

The R `vardpoor` package (Breidaks et al., 2016), created by researchers at the Central Statistical Bureau of Latvia, includes a Gini coefficient calculation using the ultimate cluster method. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the vardpoor library
library(vardpoor)

# load the laeken library
library(laeken)

# load the synthetic EU statistics on income & living conditions
data(eusilc)

# make all column names lowercase
names(eusilc) <- tolower(names(eusilc))

# add a column with the row number
dati <- data.table::data.table(IDd = 1:nrow(eusilc), eusilc)
```

```
# calculate the gini coefficient
# using the R vardpoor library
varpoord_gini_calculation <-
  varpoord(
    # analysis variable
    Y = "eqincome",

    # weights variable
    w_final = "rb050",

    # row number variable
    ID_level1 = "IDd",

    # row number variable
    ID_level2 = "IDd",

    # strata variable
    H = "db040",

    N_h = NULL ,

    # clustering variable
    PSU = "rb030",

    # data.table
    dataset = dati,

    # gini coefficient function
    type = "lingini",

    # get linearized variable
    outp_lin = TRUE

  )



# construct a survey.design
# using our recommended setup
des_eusilc <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc
```

```
  )

# immediately run the convey_prep function on it
des_eusilc <- convey_prep(des_eusilc)

# coefficients do match
varpoord_gini_calculation$all_result$value
```

```
## [1] 26.49652
```

```
coef(svygini( ~ eqincome , des_eusilc)) * 100
```

```
## eqincome
## 26.49652
```

```
# linearized variables do match
# varpoord
lin_gini_varpoord <- varpoord_gini_calculation$lin_out$lin_gini
# convey
lin_gini_convey <-
  attr(svygini( ~ eqincome , des_eusilc , linearized = TRUE) , "linearized")

# check equality
all.equal(lin_gini_varpoord , (100 * as.numeric(lin_gini_convey)))
```

```
## [1] TRUE
```

```
# variances do not match exactly
attr(svygini( ~ eqincome , des_eusilc) , 'var') * 10000
```

```
##            eqincome
## eqincome 0.03790739
```

```
varpoord_gini_calculation$all_result$var
```

```
## [1] 0.03783931
```

```
# standard errors do not match exactly
varpoord_gini_calculation$all_result$se
```

```
## [1] 0.1945233
```

```
SE(svygini( ~ eqincome , des_eusilc)) * 100
```

```
##           eqincome
## eqincome 0.1946982
```

The variance estimate is computed by using the approximation defined in 1.5, while the linearized variable z is defined by 1.4. The functions `convey::svygini` and `vardpoor::lingini` produce the same linearized variable z.

However, the measures of uncertainty do not line up, because `library(vardpoor)` defaults to an ultimate cluster method that can be replicated with an alternative setup of the `survey.design` object.

```
# within each strata, sum up the weights
cluster_sums <-
  aggregate(eusilc$rb050 , list(eusilc$db040) , sum)

# name the within-strata sums of weights the `cluster_sum`
names(cluster_sums) <- c("db040" , "cluster_sum")

# merge this column back onto the data.frame
eusilc <- merge(eusilc , cluster_sums)

# construct a survey.design
# with the fpc using the cluster sum
des_eusilc_ultimate_cluster <-
  svydesign(
    ids = ~ rb030 ,
    strata = ~ db040 ,
    weights = ~ rb050 ,
    data = eusilc ,
    fpc = ~ cluster_sum
  )

# again, immediately run the convey_prep function on the `survey.design`
des_eusilc_ultimate_cluster <-
  convey_prep(des_eusilc_ultimate_cluster)

# matches
attr(svygini( ~ eqincome , des_eusilc_ultimate_cluster) , 'var') * 10000
```

```
##            eqincome
## eqincome 0.03783931
```

```
varpoord_gini_calculation$all_result$var
```

```
## [1] 0.03783931
```

```
# matches
varpoord_gini_calculation$all_result$se
```

```
## [1] 0.1945233
```

```
SE(svygini( ~ eqincome , des_eusilc_ultimate_cluster)) * 100
```

```
##           eqincome
## eqincome 0.1945233
```

3.5 Zenga index (svyzenga)

Proposed by Zenga (2007), the Zenga index is another inequality measure related to the Lorenz curve with interesting properties. For continuous populations, it is defined as:

$$Z = 1 - \int_0^1 \frac{L(p)}{p} \cdot \frac{1-p}{1-L(p)} dp$$

In practice, `convey` uses the Barabesi et al. (2016) estimator, based on the finite population Zenga index, and a variance estimator derived using Deville's (1999) influence function method.

Alternatively, Langel and Tillé (2012) uses an estimator based on smoothed quantiles and a variance estimator based on the Demnati and Rao (2004) linearization method. However, Barabesi et al. (2016) finds that both estimators behave very similarly.

The intuition behind the Zenga index is based on argument similar to the Quintile Share Ratio or the Palma index. While the QSR and Palma indices are based on ratios of lower and upper shares, the Zenga scores (but not the measure itself!) is the ratio of lower and upper means. The Zenga index is the average of those scores.

For a given p, when the mean income of the poorest is not much smaller than the mean income of the richest, the ratio is close to one, resulting in Z(p) close to zero. However, when the lower mean is much smaller than the upper mean, the ratio goes to 0, resulting in a Z(p) close to one.

Take the mean income of the poorest 20% of a population: this is the lower mean. The "complement" is the mean income of the richest 80%: this is the upper mean. If the lower mean is $100 and the upper mean is $1,000, then the respective point on the Zenga curve would associate the fraction 20% to 1 - ($100 / $1,000) = 0.9. In other words, Z(20%) = 90%.

In this case, we took a fixed p = 20%; however, the full Zenga curve would be computed varying p from 0% to 100%. Then, the Zenga index is the integral of this curve in the interval (0,1).

Unlike the relationship between the Gini coefficient and the Lorenz curve, this allows for a more straightforward graphical interpretation (see Figure 1 from Langel and Tillé (2012)) of the Zenga index. In fact, the Zenga index is the mean of the Y axis of the Zenga coordinates.

According to Langel and Tillé (2012), "the Zenga index is significantly less affected by extreme observations than the Gini index. This is a very important advantage of the Zenga index because inference from income data is often confronted with extreme values".

A replication example

While it is not possible to reproduce the exact results because their simulation includes random sampling, we can replicate a Monte Carlo experiment similar to the one presented by Barabesi et al. (2016) in Table 2.

Load and prepare the data set:

```
library(convey)
library(survey)
library(parallel)

# create a temporary file on the local disk
tf <- tempfile()

# store the location of the zip file
data_zip <-
  "https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-

# download to the temporary file
download.file(data_zip , tf , mode = 'wb')

# unzip the contents of the archive to a temporary directory
td <- file.path(tempdir() , "shiw2012")
data_files <- unzip(tf , exdir = td)
```

```
# load the household, consumption and income from SHIW (2012) survey data
hhr.df <- read.csv(grep("carcom12" , data_files , value = TRUE))
cns.df <- read.csv(grep("risfam12" , data_files , value = TRUE))
inc.df <- read.csv(grep("rper12" , data_files , value = TRUE))

# fixes names
colnames(cns.df)[1] <- tolower(colnames(cns.df))[1]

# household count should be 8151
stopifnot(sum(!duplicated(hhr.df$nquest)) == 8151)

# person count should be 20022
stopifnot(sum(!duplicated(paste0(
  hhr.df$nquest , "-" , hhr.df$nord
))) == 20022)

# income recipients: should be 12986
stopifnot(sum(hhr.df$PERC) == 12986)

# combine household roster and income datasets
pop.df <-
  merge(
    hhr.df ,
    inc.df ,
    by = c("nquest" , "nord") ,
    all.x = TRUE ,
    sort = FALSE
  )

# compute household-level income
hinc.df <-
  pop.df[pop.df$PERC == 1 ,
         c("nquest" , "nord" , "PERC" , "Y" , "YL" , "YM" , "YT" , "YC") ,
         drop = FALSE]

hinc.df <-
  aggregate(rowSums(hinc.df[, c("YL"  , "YM" , "YT") , drop = FALSE]) ,
            list("nquest" = hinc.df$nquest) ,
            sum ,
            na.rm = TRUE)

pop.df <-
  merge(pop.df ,
        hinc.df ,
        by = "nquest" ,
```

```
        all.x = TRUE ,
        sort = FALSE)

# combine with consumption data
pop.df <-
  merge(pop.df ,
        cns.df ,
        by = "nquest" ,
        all.x = TRUE ,
        sort = FALSE)

# treat household-level sample as the finite population
pop.df <-
  pop.df[!duplicated(pop.df$nquest) , , drop = FALSE]
```

For the Monte Carlo experiment, we take `5000` samples of (expected) size `1000` using Poisson sampling[1]:

```
# set up monte carlo attributes
mc.rep <- 5000L
n.size = 1000L
N.size = nrow(pop.df)

# calculate first order probabilities
pop.df$pik <-
  n.size * (abs(pop.df$C) / sum(abs(pop.df$C)))

# calculate second order probabilities
pikl <- pop.df$pik %*% t(pop.df$pik)
diag(pikl) <- pop.df$pik

# set random number seed
set.seed(1997)

# create list of survey design objects
# using poisson sampling
survey.list <-
  lapply(seq_len(mc.rep) ,
         function(this.rep) {
           smp.ind <- which(runif(N.size) <= pop.df$pik)
           smp.df <- pop.df[smp.ind , , drop = FALSE]
           svydesign(
             ids = ~ nquest ,
```

[1] but not Conditional Poisson Sampling.

```
            data = smp.df ,
            fpc = ~ pik ,
            nest = FALSE ,
            pps = ppsmat(pikl[smp.ind , smp.ind]) ,
            variance = "HT"
          )
        })
```

... and estimate the Zenga index using each sample:

```
# estimate zenga index for each sample
zenga.estimate.list <-
  lapply(survey.list ,
         function(this.sample) {
           svyzenga( ~ x ,
                     subset(this.sample , x > 0) ,
                     na.rm = TRUE)
         })
```

Then, we evaluate the Percentage Relative Bias (PRB) of the Zenga index estimator. Under this scenario, the PRB of the Zenga index estimator is 0.3397%, a result similar to the `0.321` shown in Table 2.

```
# compute the (finite population) Zenga index parameter
theta.pop <-
  convey:::CalcZenga(pop.df$x , ifelse(pop.df$x > 0 , 1 , 0))

# estimate the expected value of the Zenga index estimator
# using the average of the estimates
theta.exp <- mean(sapply(zenga.estimate.list , coef))

# estimate the percentage relative bias
100 * (theta.exp / theta.pop - 1)
```

```
## [1] 0.3396837
```

Then, we evaluate the PRB of the variance estimator of the Zenga index estimator. Under this scenario, the PRB of the Zenga index variance estimator is -0.4954%, another result similar to the `-0.600` shown in Table 2.

```
# estimate the variance of the Zenga index estimator
# using the variance of the estimates
(vartheta.popest <- var(sapply(zenga.estimate.list , coef)))
```

```
## [1] 9.47244e-05
```

```
# estimate the expected value of the Zenga index variance estimator
# using the expected of the variance estimates
(vartheta.exp <- mean(sapply(zenga.estimate.list , vcov)))
```

```
## [1] 9.425515e-05
```

```
# estimate the percentage relative bias
100 * (vartheta.exp / vartheta.popest - 1)
```

```
## [1] -0.4953863
```

Next, we evaluate the Percentage Coverage Rate (PCR). In theory, under repeated sampling, the estimated 95% CIs should cover the population parameter 95% of the time. We can evaluate that using:

```
# estimate confidence intervals of the Zenga index
# for each of the samples
est.cis <-
  t(sapply(zenga.estimate.list, function(this.stat)
    c(confint(this.stat))))

# evaluate
prop.table(table((theta.pop > est.cis[, 1]) &
                   (theta.pop < est.cis[, 2])))
```

```
##
##  FALSE   TRUE
## 0.0548 0.9452
```

Our estimated 95% CIs cover the parameter 94.52% of the time, which is very close to the nominal rate and also similar to the `94.5` PCR shown in Table 2.

3.6 Entropy-based Measures

Entropy is a concept derived from information theory, meaning the expected amount of information given the occurrence of an event. Following (Shannon, 1948), given an event y with probability density function $f(\cdot)$, the information content given the occurrence of y can be defined as $g(f(y)) := -\log f(y)$. Therefore, the expected information or, put simply, the *entropy* is

$$H(f) \colon = -E[\log f(y)] = -\int_{-\infty}^{\infty} f(y) \log f(y) dy$$

Assuming a discrete distribution, with p_k as the probability of occurring event $k \in K$, the entropy formula takes the form:

$$H = -\sum_{k \in K} p_k \log p_k.$$

The main idea behind it is that the expected amount of information of an event is inversely proportional to the probability of its occurrence. In other words, the information derived from the observation of a rare event is higher than of the information of more probable events.

Using ideas presented in Cowell et al. (2009), substituting the density function by the income share of an individual:

$$s(q) = F^{-1}(q) / \int_0^1 F^{-1}(t) dt = y/\mu$$

the entropy function becomes the Theil (or Theil-T) inequality index:

$$I_{Theil} = \int_0^{\infty} \frac{y}{\mu} \log \left(\frac{y}{\mu} \right) dF(y) = -H(s)$$

Therefore, the entropy-based inequality measure increases as a person's income y deviates from the mean μ. This is the basic idea behind entropy-based inequality measures.

3.7 Generalized Entropy and Decomposition (svygei, svygeidec)

Using a generalization of the information function, now defined as:

$$g(f) = \frac{1}{\alpha - 1} [1 - f^{\alpha - 1}]$$

the α-class entropy is:

$$H^{(\alpha)}(f) = \frac{1}{\alpha - 1} \left[1 - \int_{-\infty}^{\infty} f(y)^{\alpha - 1} f(y) dy \right].$$

This relates to a class of inequality measures, the Generalized entropy indices, defined as:

$$GE^{(\alpha)} = \frac{1}{\alpha^2 - \alpha} \int_0^\infty \left[\left(\frac{y}{\mu} \right)^\alpha - 1 \right] dF(x) = -\frac{-H_\alpha(s)}{\alpha}.$$

The parameter α also has an economic interpretation: as α increases, the influence of high incomes upon the index increases. In some cases, this measure takes special forms, such as the mean log deviation and the aforementioned Theil index.

Biewen and Jenkins (2003) use the following finite-population as the basis for a plugin estimator:

$$GE^{(\alpha)} = \begin{cases} (\alpha^2 - \alpha)^{-1} [U_0^{\alpha-1} U_1^{-\alpha} U_\alpha - 1], & \text{if } \alpha \in \mathbb{R} \quad \{0, 1\} \\ -T_0 U_0^{-1} + \log(U_1/U_0), & \text{if } \alpha \to 0 \\ T_1 U_1^{-1} - \log(U_1/U_0), & \text{if } \alpha \to 1 \end{cases}$$

where $U_\gamma = \sum_{i \in U} y_i^\gamma$ and $T_\gamma = \sum_{i \in U} y_i^\gamma \log y_i$. Since those are all functions of totals, the linearization of the indices are easily achieved using the theorems described in Deville (1999).

This class of inequality measure also has several desirable properties, such as additive decomposition. Additive decomposition allows researchers to compare the effects of inequality within and between population groups on the population's level of inequality. Put simply, taking G groups, an additive decomposable index allows for:

$$I(\mathbf{y}) = I_{Within} + I_{Between}$$

where $I_{Within} = \sum_{g \in G} W_g I(\mathbf{y}_g)$, with W_g being measure-specific group weights; and $I_{Between}$ is a function of the group means and population sizes.

A replication example

In July 2006, Jenkins (2008) presented at the North American Stata Users' Group Meetings on the stata Generalized Entropy Index command. The example below reproduces those statistics.

Load and prepare the same data set:

```r
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the foreign library
library(foreign)

# create a temporary file on the local disk
tf <- tempfile()

# store the location of the presentation file
presentation_zip <-
  "http://repec.org/nasug2006/nasug2006_jenkins.zip"

# download jenkins' presentation to the temporary file
download.file(presentation_zip , tf , mode = 'wb')

# unzip the contents of the archive
presentation_files <- unzip(tf , exdir = tempdir())

# load the institute for fiscal studies' 1981, 1985, and 1991 data.frame objects
x81 <-
  read.dta(grep("ifs81" , presentation_files , value = TRUE))
x85 <-
  read.dta(grep("ifs85" , presentation_files , value = TRUE))
x91 <-
  read.dta(grep("ifs91" , presentation_files , value = TRUE))

# stack each of these three years of data into a single data.frame
x <- rbind(x81 , x85 , x91)
```

Replicate the author's survey design statement from stata code..

```
. * account for clustering within HHs
. version 8: svyset [pweight = wgt], psu(hrn)
pweight is wgt
psu is hrn
construct an
```

.. into R code:

```
# initiate a linearized survey design object
y <- svydesign( ~ hrn , data = x , weights = ~ wgt)

# immediately run the `convey_prep` function on the survey design
z <- convey_prep(y)
```

Replicate the author's subset statement and each of his svygei results..

```
. svygei x if year == 1981

Warning: x has 20 values = 0. Not used in calculations

Complex survey estimates of Generalized Entropy inequality indices

pweight: wgt                                      Number of obs    = 9752
Strata: <one>                                     Number of strata = 1
PSU: hrn                                          Number of PSUs   = 7459
                                                  Population size  = 54766261
------------------------------------------------------------------------------
Index    |  Estimate   Std. Err.      z      P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
GE(-1)   |  .1902062   .02474921     7.69    0.000      .1416987   .2387138
MLD      |  .1142851   .00275138    41.54    0.000      .1088925   .1196777
Theil    |  .1116923   .00226489    49.31    0.000      .1072532   .1161314
GE(2)    |   .128793   .00330774    38.94    0.000      .1223099    .135276
GE(3)    |  .1739994   .00662015    26.28    0.000      .1610242   .1869747
------------------------------------------------------------------------------
```

..using R code:

```
z81 <- subset(z , year == 1981)

svygei( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = -1)
```

```
##            gei     SE
## eybhc0 0.19021 0.0247
```

```
svygei( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 0)
```

```
##            gei     SE
## eybhc0 0.11429 0.0028
```

```
svygei( ~ eybhc0 , subset(z81 , eybhc0 > 0))
```

```
##            gei     SE
## eybhc0 0.11169 0.0023
```

```
svygei( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 2)
```

```
##            gei     SE
## eybhc0 0.12879 0.0033
```

```
svygei( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 3)
```

```
##          gei     SE
## eybhc0 0.174 0.0066
```

Confirm this replication applies for subsetted objects as well. Compare stata output..

```
. svygei x if year == 1985 & x >= 1

Complex survey estimates of Generalized Entropy inequality indices

pweight: wgt                                    Number of obs    = 8969
Strata: <one>                                   Number of strata = 1
PSU: hrn                                        Number of PSUs   = 6950
                                                Population size  = 55042871
------------------------------------------------------------------------------
Index    |  Estimate   Std. Err.      z      P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
GE(-1)   |  .1602358   .00936931    17.10    0.000      .1418723   .1785993
MLD      |   .127616   .00332187    38.42    0.000      .1211052   .1341267
Theil    |  .1337177   .00406302    32.91    0.000      .1257543    .141681
GE(2)    |  .1676393   .00730057    22.96    0.000      .1533304   .1819481
GE(3)    |  .2609507   .01850689    14.10    0.000      .2246779   .2972235
------------------------------------------------------------------------------
```

..to R code:

```
z85 <- subset(z , year == 1985)

svygei( ~ eybhc0 , subset(z85 , eybhc0 > 1) , epsilon = -1)
```

```
##             gei     SE
## eybhc0 0.16024 0.0094
```

```
svygei( ~ eybhc0 , subset(z85 , eybhc0 > 1) , epsilon = 0)
```

```
##             gei     SE
## eybhc0 0.12762 0.0033
```

```
svygei( ~ eybhc0 , subset(z85 , eybhc0 > 1))
```

```
##             gei     SE
## eybhc0 0.13372 0.0041
```

```
svygei( ~ eybhc0 , subset(z85 , eybhc0 > 1) , epsilon = 2)
```

```
##             gei     SE
## eybhc0 0.16764 0.0073
```

```
svygei( ~ eybhc0 , subset(z85 , eybhc0 > 1) , epsilon = 3)
```

```
##             gei     SE
## eybhc0 0.26095 0.0185
```

Replicate the author's decomposition by population subgroup (work status) shown on PDF page 57..

```
# define work status (PDF page 22)
z <-
  update(z , wkstatus = c(1 , 1 , 1 , 1 , 2 , 3 , 2 , 2)[as.numeric(esbu)])
z <-
  update(z , wkstatus = factor(wkstatus , labels = c("1+ ft working" , "no ft working"

# subset to 1991 and remove records with zero income
z91 <- subset(z , year == 1991 & eybhc0 > 0)

# population share
svymean( ~ wkstatus, z91)
```

```
##                            mean     SE
## wkstatus1+ ft working   0.61724 0.0067
## wkstatusno ft working   0.20607 0.0059
## wkstatuselderly         0.17669 0.0046
```

```
# mean
svyby( ~ eybhc0, ~ wkstatus, z91, svymean)
```

```
##                     wkstatus   eybhc0       se
## 1+ ft working 1+ ft working 278.8040 3.703790
## no ft working no ft working 151.6317 3.153968
## elderly             elderly 176.6045 4.661740
```

```
# subgroup indices: ge_k
svyby( ~ eybhc0 , ~ wkstatus , z91 , svygei , epsilon = -1)
```

```
##                     wkstatus     eybhc0          se
## 1+ ft working 1+ ft working  0.2300708  0.02853959
## no ft working no ft working 10.9231761 10.65482557
## elderly             elderly  0.1932164  0.02571991
```

```
svyby( ~ eybhc0 , ~ wkstatus , z91 , svygei , epsilon = 0)
```

```
##                     wkstatus    eybhc0          se
## 1+ ft working 1+ ft working 0.1536921 0.006955506
## no ft working no ft working 0.1836835 0.014740510
## elderly             elderly 0.1653658 0.016409770
```

```
svyby( ~ eybhc0 , ~ wkstatus , z91 , svygei , epsilon = 1)
```

```
##                     wkstatus    eybhc0          se
## 1+ ft working 1+ ft working 0.1598558 0.008327994
## no ft working no ft working 0.1889909 0.016766120
## elderly             elderly 0.2023862 0.027787224
```

```
svyby( ~ eybhc0 , ~ wkstatus , z91 , svygei , epsilon = 2)
```

```
##                     wkstatus    eybhc0         se
## 1+ ft working 1+ ft working 0.2130664 0.01546521
## no ft working no ft working 0.2846345 0.06016394
## elderly             elderly 0.3465088 0.07362898
```

```
# GE decomposition
svygeidec( ~ eybhc0, ~ wkstatus, z91, epsilon = -1)
```

```
##         gei decomposition     SE
## total            3.682893 3.3999
## within           3.646572 3.3998
## between          0.036321 0.0028
```

```
svygeidec( ~ eybhc0, ~ wkstatus, z91, epsilon = 0)
```

```
##         gei decomposition     SE
## total            0.195236 0.0065
## within           0.161935 0.0061
## between          0.033301 0.0025
```

```
svygeidec( ~ eybhc0, ~ wkstatus, z91, epsilon = 1)
```

```
##         gei decomposition     SE
## total            0.200390 0.0079
## within           0.169396 0.0076
## between          0.030994 0.0022
```

```
svygeidec( ~ eybhc0, ~ wkstatus, z91, epsilon = 2)
```

```
##         gei decomposition     SE
## total            0.274325 0.0167
## within           0.245067 0.0164
## between          0.029258 0.0021
```

For additional usage examples of `svygei` or `svygeidec`, type `?convey::svygei` or `?convey::svygeidec` in the R console.

3.8 J-Divergence and Decomposition (svyjdiv, svyjdivdec)

The J-divergence measure (Rohde, 2016) can be seen as the sum of GE_0 and GE_1, satisfying axioms that, individually, those two indices do not. The J-divergence measure is defined as:

$$J = \frac{1}{N} \sum_{i \in U} \left(\frac{y_i}{\mu} - 1 \right) \ln \left(\frac{y_i}{\mu} \right)$$

Since it is a sum of two additive decomposable measures, J itself is decomposable. The within and between parts of the J-divergence decomposition are:

$$J_B = \sum_{g \in G} \frac{N_g}{N} \left(\frac{\mu_g}{\mu} - 1 \right) \ln \left(\frac{\mu_g}{\mu} \right)$$

$$J_W = \sum_{g \in G} \left[\frac{Y_g}{Y} GE_g^{(1)} + \frac{N_g}{N} GE_g^{(0)} \right] = \sum_{g \in G} \left[\left(\frac{Y_g N + Y N_g}{YN} \right) J_g \right] - \sum_{g \in G} \left[\frac{Y_g}{Y} GE_g^{(1)} + \frac{N_g}{N} GE_g^{(0)} \right]$$

where the subscript g denotes one of the G subgroups in the population. The main difference with the additively decomposable Generalized Entropy family is that the "within" component of the J-divergence index cannot be seen as a weighted contribution of the J-divergence across groups. So J_W cannot be interpreted as a weighted mean of the J-divergence across groups, but as the sum of two terms: (1) the income-share weighted mean of the $GE^{(1)}$ across groups; and (2) the population-share weighted mean of the $GE^{(0)}$ across groups. In other words, the "within" component still accounts for the inequality in each group; it just does not have a direct interpretation.

A replication example

First, we should check that the finite-population values make sense. The J-divergence can be seen as the sum of $GE^{(0)}$ and $GE^{(1)}$. So, taking the starting population from the `svyzenga` section of this text, we have:

```
# compute finite population J-divergence
(jdivt.pop <-
   (convey:::CalcJDiv(pop.df$x , ifelse(pop.df$x > 0 , 1 , 0))))
```

```
## [1] 0.4332649
```

```
# compute finite population GE indices
(gei0.pop <-
    convey:::CalcGEI(pop.df$x , ifelse(pop.df$x > 0 , 1 , 0) , 0))
```

```
## [1] 0.2215037
```

```
(gei1.pop <-
    convey:::CalcGEI(pop.df$x , ifelse(pop.df$x > 0 , 1 , 0) , 1))
```

```
## [1] 0.2117612
```

```
# check equality; should be true
all.equal(jdivt.pop , gei0.pop + gei1.pop)
```

```
## [1] TRUE
```

As expected, the J-divergence matches the sum of GEs in the finite population. And as we've checked the GE measures before, the J-divergence computation function seems safe.

In order to assess the estimators implemented in `svyjdiv` and `svyjdivdec`, we can run a Monte Carlo experiment. Using the same samples we used in the `svyzenga` replication example, we have:

```
# estimate J-divergence with each sample
jdiv.estimate.list <-
  lapply(survey.list ,
         function(this.sample) {
           svyjdiv( ~ x ,
                    subset(this.sample , x > 0) ,
                    na.rm = TRUE)
         })

# compute the (finite population overall) J-divergence
(theta.pop <-
    convey:::CalcJDiv(pop.df$x , ifelse(pop.df$x > 0 , 1 , 0)))
```

```
## [1] 0.4332649
```

```
# estimate the expected value of the J-divergence estimator
# using the average of the estimates
(theta.exp <- mean(sapply(jdiv.estimate.list , coef)))
```

```
## [1] 0.4327823
```

```
# estimate the percentage relative bias
100 * (theta.exp / theta.pop - 1)
```

```
## [1] -0.1113765
```

```
# estimate the variance of the J-divergence estimator
# using the variance of the estimates
(vartheta.popest <- var(sapply(jdiv.estimate.list , coef)))
```

```
## [1] 0.0005434848
```

```
# estimate the expected value of the J-divergence index variance estimator
# using the expected of the variance estimates
(vartheta.exp <- mean(sapply(jdiv.estimate.list , vcov)))
```

```
## [1] 0.0005342947
```

```
# estimate the percentage relative bias of the variance estimator
100 * (vartheta.exp / vartheta.popest - 1)
```

```
## [1] -1.690964
```

For the decomposition, we repeat the same procedure:

```
# estimate J-divergence decomposition with each sample
jdivdec.estimate.list <-
  lapply(survey.list ,
         function(this.sample) {
           svyjdivdec( ~ x ,
                       ~ SEX ,
                       subset(this.sample , x > 0) ,
                       na.rm = TRUE)
         })

# compute the (finite population) J-divergence decomposition per sex
jdivt.pop <-
  convey:::CalcJDiv(pop.df$x , ifelse(pop.df$x > 0 , 1 , 0))

overall.mean <- mean(pop.df$x[pop.df$x > 0])

group.mean <-
  c(by(pop.df$x[pop.df$x > 0] , list("SEX" = factor(pop.df$SEX[pop.df$x > 0])) , FUN = mean))

group.pshare <-
  c(prop.table(by(rep(1 , sum(
    pop.df$x > 0
  )) , list(
    "SEX" = factor(pop.df$SEX[pop.df$x > 0])
  ) , FUN = sum)))

jdivb.pop <-
  sum(group.pshare * (group.mean / overall.mean - 1) * log(group.mean / overall.mean))
```

```
jdivw.pop <- jdivt.pop - jdivb.pop

(theta.pop <- c(jdivt.pop , jdivw.pop , jdivb.pop))
```

```
## [1] 0.433264877 0.428398928 0.004865949
```

```
# estimate the expected value of the J-divergence decomposition estimator
# using the average of the estimates
(theta.exp <- rowMeans(sapply(jdivdec.estimate.list , coef)))
```

```
##       total      within     between
## 0.432782322 0.427480257 0.005302065
```

```
# estimate the percentage relative bias
100 * (theta.exp / theta.pop - 1)
```

```
##      total     within    between
## -0.1113765 -0.2144430  8.9626164
```

The estimated PRB for the total is the same as before, so we will focus on the within and between components. While the within component has a small relative bias (-0.21%), the between component PRB is significant, amounting to 8.96%.

For the variance estimator, we do:

```
# estimate the variance of the J-divergence estimator
# using the variance of the estimates
(vartheta.popest <-
   diag(var(t(
     sapply(jdivdec.estimate.list , coef)
   ))))
```

```
##        total       within      between
## 5.434848e-04 5.391901e-04 8.750879e-06
```

```
# estimate the expected value of the J-divergence index variance estimator
# using the expected of the variance estimates
(vartheta.exp <-
    rowMeans(sapply(jdivdec.estimate.list , function(z)
      diag(vcov(
        z
      )))))
```

```
##        total       within      between
## 5.342947e-04 5.286750e-04 8.891772e-06
```

```
# estimate the percentage relative bias of the variance estimator
100 * (vartheta.exp / vartheta.popest - 1)
```

```
##     total    within   between
## -1.690964 -1.950177  1.610034
```

The PRB of the variance estimators for both components are small: -1.95% for the within component and 1.61% for the between component.

Now, how much should we care about the between component bias? Our simulations show that the Squared Bias of this estimator accounts for less than 2% of its Mean Squared Error:

```
theta.bias2 <- (theta.exp - theta.pop) ^ 2
theta.mse <- theta.bias2 + vartheta.popest
100 * (theta.bias2 / theta.mse)
```

```
##      total     within    between
## 0.04282728 0.15627854 2.12723212
```

Next, we evaluate the Percentage Coverage Rate (PCR). In theory, under repeated sampling, the estimated 95% CIs should cover the population parameter 95% of the time. We can evaluate that using:

```
# estimate confidence intervals of the Zenga index
# for each of the samples
est.coverage <-
  t(sapply(jdivdec.estimate.list, function(this.stat)
    confint(this.stat)[, 1] <= theta.pop &
      confint(this.stat)[, 2] >= theta.pop))

# evaluate
colMeans(est.coverage)
```

```
##   total  within between
##  0.9390  0.9376  0.9108
```

Our coverages are not too far from the nominal coverage level of 95%, however the bias of the between component estimator can affect its coverage rate.

For additional usage examples of `svyjdiv` or `svyjdivdec`, type `?convey::svyjdiv` or `?convey::svyjdivdec` in the R console.

3.9 Atkinson index (svyatk)

Although the original formula was proposed in Atkinson (1970), the estimator used here comes from Biewen and Jenkins (2003):

$$\widehat{A}_\epsilon = \begin{cases} 1 - \widehat{U}_0^{-\epsilon/(1-\epsilon)} \widehat{U}_1^{-1} \widehat{U}_{1-\epsilon}^{1/(1-\epsilon)}, & \text{if } \epsilon \in \mathbb{R}_+ \quad \{1\} \\ 1 - \widehat{U}_0 \widehat{U}_0^{-1} exp(\widehat{T}_0 \widehat{U}_0^{-1}), & \text{if } \epsilon \to 1 \end{cases}$$

The ϵ is an inequality aversion parameter: as it approaches infinity, more weight is given to incomes in bottom of the distribution.

A replication example

In July 2006, Jenkins (2008) presented at the North American Stata Users' Group Meetings on the stata Atkinson Index command. The example below reproduces those statistics.

Load and prepare the same data set:

```
# load the convey package
library(convey)

# load the survey library
library(survey)

# load the foreign library
library(foreign)

# create a temporary file on the local disk
tf <- tempfile()

# store the location of the presentation file
presentation_zip <-
  "http://repec.org/nasug2006/nasug2006_jenkins.zip"

# download jenkins' presentation to the temporary file
download.file(presentation_zip , tf , mode = 'wb')

# unzip the contents of the archive
presentation_files <- unzip(tf , exdir = tempdir())

# load the institute for fiscal studies' 1981, 1985, and 1991 data.frame objects
x81 <-
```

```
  read.dta(grep("ifs81" , presentation_files , value = TRUE))
x85 <-
  read.dta(grep("ifs85" , presentation_files , value = TRUE))
x91 <-
  read.dta(grep("ifs91" , presentation_files , value = TRUE))

# stack each of these three years of data into a single data.frame
x <- rbind(x81 , x85 , x91)
```

Replicate the author's survey design statement from stata code..

```
. * account for clustering within HHs
. version 8: svyset [pweight = wgt], psu(hrn)
pweight is wgt
psu is hrn
construct an
```

.. into R code:

```
# initiate a linearized survey design object
y <- svydesign( ~ hrn , data = x , weights = ~ wgt)

# immediately run the `convey_prep` function on the survey design
z <- convey_prep(y)
```

Replicate the author's subset statement and each of his svyatk results with stata..

```
. svyatk x if year == 1981

Warning: x has 20 values = 0. Not used in calculations

Complex survey estimates of Atkinson inequality indices

pweight: wgt                                    Number of obs    = 9752
Strata: <one>                                   Number of strata = 1
PSU: hrn                                        Number of PSUs   = 7459
                                                Population size  = 54766261
------------------------------------------------------------------------------
Index    |  Estimate   Std. Err.      z      P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
A(0.5)   |  .0543239   .00107583    50.49    0.000      .0522153   .0564324
A(1)     |  .1079964   .00245424    44.00    0.000      .1031862   .1128066
A(1.5)   |  .1701794   .0066943    25.42    0.000      .1570588      .1833
```

```
A(2)     |  .2755788   .02597608    10.61    0.000      .2246666    .326491
A(2.5)   |  .4992701   .06754311     7.39    0.000       .366888   .6316522
------------------------------------------------------------------------------
```

..using R code:

```
z81 <- subset(z , year == 1981)

svyatk( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 0.5)
```

```
##         atkinson     SE
## eybhc0 0.054324 0.0011
```

```
svyatk( ~ eybhc0 , subset(z81 , eybhc0 > 0))
```

```
##         atkinson     SE
## eybhc0    0.108 0.0025
```

```
svyatk( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 1.5)
```

```
##         atkinson     SE
## eybhc0  0.17018 0.0067
```

```
svyatk( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 2)
```

```
##         atkinson    SE
## eybhc0  0.27558 0.026
```

```
svyatk( ~ eybhc0 , subset(z81 , eybhc0 > 0) , epsilon = 2.5)
```

```
##         atkinson     SE
## eybhc0  0.49927 0.0675
```

Confirm this replication applies for subsetted objects as well, comparing stata code..

```
. svyatk x if year == 1981 & x >= 1

Complex survey estimates of Atkinson inequality indices

pweight: wgt                                      Number of obs    = 9748
```

```
Strata: <one>                                     Number of strata = 1
PSU: hrn                                          Number of PSUs   = 7457
                                                  Population size  = 54744234
------------------------------------------------------------------------------
Index    |  Estimate   Std. Err.      z      P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
A(0.5)   |  .0540059   .00105011    51.43    0.000      .0519477   .0560641
A(1)     |  .1066082   .00223318    47.74    0.000      .1022313   .1109852
A(1.5)   |  .1638299   .00483069    33.91    0.000       .154362   .1732979
A(2)     |  .2443206   .01425258    17.14    0.000      .2163861   .2722552
A(2.5)   |   .394787   .04155221     9.50    0.000      .3133461   .4762278
------------------------------------------------------------------------------
```

..to R code:

```
z81_two <- subset(z , year == 1981 & eybhc0 > 1)

svyatk( ~ eybhc0 , z81_two , epsilon = 0.5)
```

```
##        atkinson     SE
## eybhc0 0.054006 0.0011
```

```
svyatk( ~ eybhc0 , z81_two)
```

```
##        atkinson     SE
## eybhc0  0.10661 0.0022
```

```
svyatk( ~ eybhc0 , z81_two , epsilon = 1.5)
```

```
##        atkinson     SE
## eybhc0  0.16383 0.0048
```

```
svyatk( ~ eybhc0 , z81_two , epsilon = 2)
```

```
##        atkinson     SE
## eybhc0  0.24432 0.0143
```

```
svyatk( ~ eybhc0 , z81_two , epsilon = 2.5)
```

```
##        atkinson     SE
## eybhc0  0.39479 0.0416
```

For additional usage examples of `svyatk`, type `?convey::svyatk` in the R console.

3.10 Which inequality measure should be used?

The variety of inequality measures begs a question: which inequality measure should be used? In fact, this is a very important question. However, the nature of it is not statistical or mathematical, but ethical. This section aims to clarify and, while not proposing a "perfect measure", to provide the reader with some initial guidance about which measure to use.

The most general way to analyze if one distribution is more equally distributed than another is by the Lorenz curve. When $L_A(p) \geqslant L_B(p), \forall p \in [0,1]$, it is said that A is more equally distributed than B. Technically, we say that *A (Lorenz) dominates B*. (Krämer (1998) and Mosler (1994) provide helpful insights to how majorization, Lorenz dominance, and inequality measurement are connected. On the topic of majorization, Hardy et al. (1934) is still the main reference, while Marshall et al. (2011) provide a more modern approach.) In this case, all inequality measures that satisfy basic properties[2] will agree that A is more equally distributed than B.

When this dominance fails, i.e., when Lorenz curves do cross, Lorenz ordering is impossible. Then, under such circumstances, the choice of which inequality measure to use becomes relevant.

Each inequality measure is a result of a subjective understanding of what is a fair distribution. As Dalton (1920, p.348) puts it, "the economist is primarily interested, not in the distribution of income as such, but in the effects of the distribution of income upon the distribution and total amount of economic welfare, which may be derived from income." The importance of how economic welfare is defined is once again expressed by Atkinson (1970), where an inequality measure is direclty derived from a class of welfare functions. Even when a welfare function is not explicit, such as in the Gini index, we must agree that an implicit, subjective judgement of the impact of inequality on social welfare is assumed.

The idea of what is a fair distribution is a matter of Ethics, a discipline within the realm of Philosophy. Yet, as Fleurbaey (1996, Ch.1) proposes, the analyst should match socially supported moral values and theories of justice to the set of technical tools for policy evaluation.

Although this can be a useful principle, a more objective answer is needed. By knowing the nature and properties of inequality measures, the analyst can further reduce the set of applicable inequality measures. For instance, choosing from the properties listed in Cowell (2011, p.74), if we require group-decomposability, scale invariance, population invariance, and that the estimate falls within $[0,1]$, we must resort to the Atkinson index.

Even though the discussion can go deep in technical and philosophical aspects, this choice also depends on the public. For example, it would not be surprising

[2] Namely, Schur-convexity, population invariance, and scale invariance.

if a public official has not encountered the Atkinson index; however, they might be familiar with the Gini coefficient. The same goes for publications: journalists have been introduced to the Gini index and can find it easier to compare and, therefore, write about. Also, we must admit that the Gini index is much more straightforward than any other measure.

In the end, the choice is mostly subjective and there is no consensus of which method offers the ideal inequality measure. We must remember that this choice is only problematic if Lorenz curves cross and, in that case, it is often straightforward to choose among the inequality measures available.

Chapter 4

Richness Measures

Richness (or Affluence) measures provide another approach for understanding income concentration. Unlike inequality measures, that are sensitive to the changes over the entire income distribution, richness measures are restricted to the top.

In that sense, they work like "inverted" poverty measures: while the poverty focus axiom states that poverty measures should be insensitive to changes in the incomes of the non-poor, the richness focus axiom states that richness measures should remain unaltered by changes in the incomes of the non-rich.

Also like poverty measures, richness measures also rely on a "richness threshold": those above that threshold are regarded as rich, while those below are regarded as non-rich.[1] Like for poverty measurement, these richness thresholds can be somewhat arbitrary; Medeiros (2006) provide a nice review of proposed richness thresholds and also propose one possible approach with practical and theoretical grounds.

Peichl et al. (2010) presented an axiomatic study of richness measures. They classify richness measures in two types depending on the effect of income transfers among the rich. A *concave* measure follows the concave transfer axiom (T1): the measure should increase when a progressive transfer occurs among two rich individuals. On the other hand, a *convex* measure follows the convex transfer axiom (T2): the measure should *decrease* when a progressive transfer occurs among two rich individuals.

The reasoning behind the concave transfer axiom is that such progressive transfer would make the income distribution among the rich more internally homogeneous. By clustering the incomes and assets of the rich away from the rest of society, the distribution of income becomes more polarized. On the other hand,

[1] Not necessarily, but including the poor

the convex transfer axiom means that a progressive transfer among rich individuals reduces inequality among them, so the convex inequality measure should register a reduction. In a sense, concave measures are related to the income polarization approach, meaning that increases in income polarization should increase the value of the richness measure, while convex measures are related to the income inequality approach, meaning that a reduction in inequality among the rich should also result in a reduction of a convex richness measure.

Peichl et al. (2010) also presented a general class of richness measures, proposing three particular sub-classes: the (concave) Chakravarty class, the concave-FGT class and the convex-FGT class, defined at the population level as:

$$R_{\gamma}^{Cha} = \frac{1}{N}\sum_{k\in U}\left[(1-\left(\frac{z_r}{y_k}\right)^{\gamma}\right]\delta(y_k \geq z_r), \quad \gamma > 0$$

$$R_{\gamma}^{FGT,T1} = \frac{1}{N}\sum_{k\in U}\left(\frac{y_k - z_r}{y_k}\right)^{\gamma}\delta(y_k \geq z_r), \quad \gamma \in [0,1)$$

$$R_{\gamma}^{FGT,T2} = \frac{1}{N}\sum_{k\in U}\left(\frac{y_k - z_r}{z_r}\right)^{\gamma}\delta(y_k \geq z_r), \quad \gamma > 1$$

where z_r is the richness threshold and γ is a sensitivity parameter.

To estimate these measures, Brzezinski (2014) proposed the estimators

$$\widehat{R}_{\gamma}^{Cha} = \frac{1}{\widehat{N}}\sum_{k\in s} w_k\left[(1-\left(\frac{z_r}{y_k}\right)^{\gamma}\right]\delta(y_k \geq z_r), \quad \gamma > 0$$

$$\widehat{R}_{\gamma}^{FGT,T1} = \frac{1}{\widehat{N}}\sum_{k\in s} w_k\left(\frac{y_k - z_r}{y_k}\right)^{\gamma}\delta(y_k \geq z_r), \quad \gamma \in [0,1)$$

$$\widehat{R}_{\gamma}^{FGT,T2} = \frac{1}{\widehat{N}}\sum_{k\in s} w_k\left(\frac{y_k - z_r}{z_r}\right)^{\gamma}\delta(y_k \geq z_r), \quad \gamma > 1$$

where w_k is the sampling (or calibration) weight.

In order to estimate the variance of these estimators, Brzezinski (2014) derived influence functions under the Deville (1999) approach. These functions are the ones used in the `svyrich` function.

Brzezinski (2014) also studied the reliability of the inference based on these estimators using a model-based Monte Carlo simulation, which are also valid for (design-based) simple random sampling with replacement. His results showed that inference for convex measures are unreliable for convex measures, the estimator generally produces bad results. It is highly sensitive to outliers, and confidence intervals are invalid (actual coverage is much smaller than 95%). The vignette below shows for a design-based approach with the same results.

A replication example

Brzezinski (2014) presented results using a model-based Monte Carlo — i.e., he simulated several samples from the model and compared the behavior of the estimators with the model-based parameters. In the simulation below, we take a design-based approach: we take several samples from a fixed finite population using a particular sampling design, compute the estimator on each and compare them to the finite population parameter (not the model-based parameter!).

For the sake of similarity, we start by simulating a large finite population ($N = 10^5$) using the same distribution from Brzezinski (2014):

```
# load libraries
library(survey)
library(convey)
library(sampling)
library(VGAM)

# set random seed
set.seed(2023)

# superpopulation parameters
scale.x0 <- 1
shape.theta <- 2
cha.beta <- 2
fgt.alpha <- 1.5
n.pop <- as.integer(10 ^ 5)

# generate finite population
pop.df <-
  data.frame(y1 = rparetoI(n.pop  ,  scale.x0 , shape.theta))
```

Then, we compute the finite population parameters using the simulated population:

```
# richness measures: finite population parameters
cha.scores <-
  function(y , b , rho)
    ifelse(y > rho , (1 - (rho / y) ^ b) , 0)
fgtt2.scores <-
  function(y , g , rho)
    ifelse(y > rho , (y / rho - 1) ^ a , 0)
median.fp <- quantile(pop.df$y1 , .50)
rho.fp <- 3 * median.fp
rHC.fp <- mean(pop.df$y1 > rho.fp)
rCha.fp <- mean(cha.scores(pop.df$y1  , cha.beta , rho.fp))
rFGTT2.fp <- mean(cha.scores(pop.df$y1  , fgt.alpha , rho.fp))
```

For our sampling design, we select $n = 1000$ units using multinomial sampling, with the variable `x.aux` as the size variable for the inclusion probabilities:

```
# define sample size
n.sample <- 1000L

# inclusion probability
pop.df$x.aux <- plogis( pop.df$y1 ) / 1.1
pop.df$pi1 <- sampling::inclusionprobabilities( pop.df$x.aux , n.sample )
```

We run the procedure N and store the estimate objects using the code below:

```
# define the number of simulation runs
mc.reps <- 5000L

# simulation runs
rep.list <- lapply(seq_len(mc.reps) , function(this.iter) {
  # multinomial sampling
  this.sample <- sampling::UPmultinomial(pop.df$pi1)
  this.sample <- rep(1:n.pop , this.sample)
  sample.df <- pop.df[this.sample ,]
  sample.df$weights <- 1 / sample.df$pi1
  des.obj <-
    svydesign(
      ids = ~ 1 ,
      weights = ~ weights ,
      data = sample.df ,
      nest = FALSE
    )

  # run estimation
  des.obj <- convey_prep(des.obj)
  rCha.hat <-
    svyrich(
      ~ y1 ,
      des.obj ,
      type_measure = "Cha" ,
      g = cha.beta ,
      type_thresh = "relq" ,
      percent = 3
    )
  suppressWarnings(
    rHC.hat <-
      svyrich(
        ~ y1 ,
```

```
        des.obj ,
        type_measure = "FGTT1" ,
        g = 0 ,
        type_thresh = "relq"  ,
        percent = 3
      )
  )
  suppressWarnings(
    rFGTT2.hat <-
      svyrich(
        ~ y1 ,
        des.obj ,
        type_measure = "FGTT2" ,
        g = fgt.alpha ,
        type_thresh = "relq"  ,
        percent = 3
      )
  )
  est.list <- list(rHC.hat , rCha.hat , rFGTT2.hat)
  est.list

})
```

To study the behavior of the estimators, we estimate their expected values, empirical variance (for the main parameter) and mean squared error. To study the validity of the normal approximation, we also estimate the percent coverage rate of the nominal 95% confidence interval. This is done using the function below:

```
sim.compile <- function(ll ,
                        pv ,
                        level = .95 ,
                        na.rm = FALSE) {
  # collect estimates
  mhat.vec <- sapply(ll , coef)
  vhat.vec <- sapply(ll , vcov)

  # estimate expected value
  mhat.exp <- mean(mhat.vec , na.rm = na.rm)
  vhat.exp <- mean(vhat.vec , na.rm = na.rm)

  # calculate empirical variance
  mhat.empvar <- var(mhat.vec , na.rm = na.rm)

  # estimate squared bias
```

```
  mhat.bias2 <- (mhat.exp - pv) ^ 2
  vhat.bias2 <- (vhat.exp - mhat.empvar) ^ 2

  # estimate mse
  mhat.mse <- mhat.bias2 + mhat.empvar

  # estimate coverage rate
  ci.hats <- t(sapply(ll , confint))
  ci.check <-
    matrix(as.logical(NA) , nrow = nrow(ci.hats) , ncol = 3)
  ci.check[, 1] <- ci.hats[, 1] <= pv
  ci.check[, 2] <- ci.hats[, 2] >= pv
  ci.check[, 3] <- apply(ci.check[, 1:2] , 1 , all)
  pcr.emp <- mean(ci.check[, 3] , na.rm = na.rm)

  # setup final table
  data.frame(
    "mc.reps" = length(ll) ,
    "theta" = pv ,
    "theta.hat" = mhat.exp ,
    "theta.bias2" = mhat.bias2 ,
    "theta.empvar" = mhat.empvar ,
    "theta.hat.mse" = mhat.mse ,
    "theta.varhat" = vhat.exp ,
    "pcr" = pcr.emp
  )
}
```

For the Headcount Richness Ratio (computed using the concave FGT measure), we have:

```
rhc.list <- lapply(rep.list , `[[` , 1)
sim.compile(rhc.list, rHC.fp)
```

```
##   mc.reps   theta  theta.hat  theta.bias2 theta.empvar theta.hat.mse
## 1    5000 0.05469 0.05452618 2.683797e-08 4.112616e-05    4.1153e-05
##   theta.varhat    pcr
## 1 4.336252e-05 0.9524
```

```
stopifnot(round(
  sim.compile(rhc.list, rHC.fp)["theta.bias2"] / sim.compile(rhc.list, rHC.fp)["theta.
  4
) == 0.0007)
```

```
stopifnot(round(
  sim.compile(rhc.list, rHC.fp)["theta.varhat"] / sim.compile(rhc.list, rHC.fp)["theta.empvar"] ,
  2
) == 1.05)

stopifnot(round(sim.compile(rhc.list, rHC.fp)["pcr"] , 4) == 0.9524)
```

Under this approach, the squared bias accounts for approx. 0.07% of the MSE, indicating that the MSE of the estimator is reasonably approximated by its variance. Additionally, the ratio between the expected value of the variance estimator and the empirical variance is approx. 1.05, indicating that the variance estimates are expected to be a (slighlty conservative, but) good approximation of the empirical variance. Finally, the estimated percent coverage rate of 95.24% is close to the nominal level of 95%, indicating that the confidence intervals are approximately valid.

For the Chakravarty measure, we have:

```
rcha.list <- lapply(rep.list , `[[` , 2)
sim.compile(rcha.list, rCha.fp)
```

```
##   mc.reps      theta  theta.hat  theta.bias2 theta.empvar theta.hat.mse
## 1    5000 0.02743915 0.02737177 4.538966e-09 1.428639e-05  1.429093e-05
##   theta.varhat    pcr
## 1 1.402416e-05 0.9428
```

```
stopifnot(round(
  sim.compile(rcha.list, rCha.fp)["theta.bias2"] / sim.compile(rcha.list, rCha.fp)["theta.hat.mse
  4
) == 0.0003)

stopifnot(round(
  sim.compile(rcha.list, rCha.fp)["theta.varhat"] / sim.compile(rcha.list, rCha.fp)["theta.empvar
  2
) == 0.98)

stopifnot(round(sim.compile(rcha.list, rCha.fp)["pcr"] , 4) == 0.9428)
```

Under this approach, the squared bias is approx 0.03% of the MSE, indicating that the MSE of the estimator is reasonably approximated by its variance. Additionally, the ratio between the expected value of the variance estimator and the empirical variance is approx. 0.98, indicating that the variance estimates are expected to be a good approximation of the empirical variance. Finally, the estimated percent coverage rate of 94.28% is close to the nominal level of 95%, indicating that the confidence intervals are approximately valid.

For the convex FGT measure, we have:

```
rfgtt2.list <- lapply(rep.list , `[[` , 3)
sim.compile(rfgtt2.list, rFGTT2.fp)
```

```
##   mc.reps      theta theta.hat theta.bias2 theta.empvar theta.hat.mse
## 1    5000 0.02349189 0.1045973  0.00657808   0.01865219    0.02523027
##   theta.varhat    pcr
## 1   0.01857144 0.5212
```

```
stopifnot(round(
  sim.compile(rfgtt2.list, rFGTT2.fp)["theta.bias2"] / sim.compile(rfgtt2.list, rFGTT2
  2
) == 0.26)

stopifnot(round(
  sim.compile(rfgtt2.list, rFGTT2.fp)["theta.varhat"] / sim.compile(rfgtt2.list, rFGTT
  2
) == 1)

stopifnot(round(sim.compile(rfgtt2.list, rFGTT2.fp)["pcr"] , 4) == 0.5212)
```

Under this approach, the squared bias is approx 26% of the MSE, indicating that the *bias is substantial*. The ratio between the expected value of the variance estimator and the empirical variance is approx. 1.00, indicating that the variance estimates are expected to be a good approximation of the empirical variance (but not of the MSE!). Finally, the estimated percent coverage rate of 52.12% is far from the nominal level of 95%, indicating that the confidence intervals are invalid. This comes from the fact that the estimator is very sensitive to the extreme values and is the reason why Brzezinski (2014) does not recommend using convex richness measures.

Bibliography

Alfons, A., Holzer, J., & Templ, M. (2014). *Laeken: Estimation of indicators on social exclusion and poverty* [R package version 0.4.6]. https://CRAN.R-project.org/package=laeken

Aristondo, O., De La Vega, C. L., & Urrutia, A. (2010). A new multiplicative decomposition for the foster–greer–thorbecke poverty indices. *Bulletin of Economic Research*, *62*(3), 259–267. https://doi.org/10.1111/j.1467-8586.2009.00320.x

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*(3), 244–263. https://ideas.repec.org/a/eee/jetheo/v2y1970i3p244-263.html

Barabesi, L., Diana, G., & Perri, P. F. (2016). Linearization of inequality indices in the design-based framework. *Statistics*, *50*(5), 1161–1172. https://doi.org/10.1080/02331888.2015.1135924

Bedi, T., Coudouel, A., & Simler, K. (2007). *More than a pretty picture: Using poverty maps to design better policies and interventions*. World Bank Publications.

Berger, Y. G., & Skinner, C. J. (2003). Variance estimation for a low income proportion. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(4), 457–468. https://doi.org/10.1111/1467-9876.00417

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, *137*. https://doi.org/10.1016/j.jeconom.2005.09.003

Biewen, M., & Jenkins, S. (2003). *Estimation of generalized entropy and atkinson inequality indices from complex survey data* (Discussion Papers of DIW Berlin No. 345). DIW Berlin, German Institute for Economic Research. http://EconPapers.repec.org/RePEc:diw:diwwpp:dp345

Blackburn, M. L. (1989). Poverty measurement: An index related to a theil measure of inequality. *Journal of Business & Economic Statistics*, *7*(4), 475–481. https://doi.org/10.1080/07350015.1989.10509760

Breidaks, J., Liberts, M., & Ivanova, S. (2016). Vardpoor: Estimation of indicators on social exclusion and poverty and its linearization, variance estimation [R package version 0.8.0]. *CSB*.

Breidaks, J., Liberts, M., & Ivanova, S. (2020). *vardpoor: Estimation of indicators on social exclusion and poverty and its linearization, variance*

estimation [R package version 0.20.0]. Riga, Latvia. https://csblatvia.github.io/vardpoor/

Brzezinski, M. (2014). Statistical inference for richness measures. *Applied Economics*, *46*(14), 1599–1608. https://doi.org/10.1080/00036846.2014.880106

Cobham, A., Schlogl, L., & Sumner, A. (2015, July). *Inequality and the Tails: The Palma Proposition and Ratio Revisited* (Working Papers No. 143). United Nations, Department of Economics and Social Affairs. http://www.un.org/esa/desa/papers/2015/wp143_2015.pdf

Cowell, F. A., Flachaire, E., & Bandyopadhyay, S. (2009, August). *Goodness-of-fit: An economic approach* (Economics Series Working Papers No. 444). University of Oxford, Department of Economics. https://ideas.repec.org/p/oxf/wpaper/444.html

Cowell, F. A. (2011). *Measuring inequality* (3rd ed.). Oxford University Press.

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal*, *30*. https://doi.org/10.2307/2223525

Deaton, A. (1997). *The analysis of household surveys: A microeconomic approach to development policy*. World Bank Publications.

Demnati, A., & Rao, J. N. K. (2004). Linearization Variance Estimators for Survey Data. *Survey Methodology*, *30*(1), 17–26. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004001/article/6991-eng.pdf

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, *25*(2), 193–203. http://www.statcan.gc.ca/pub/12-001-x/1999002/article/4882-eng.pdf

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, *71*. https://doi.org/10.1111/1468-0262.00399

Fleurbaey, M. (1996). *Théories économiques de la justice*. Economica.

Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, *52*(3), 761–766. http://www.jstor.org/stable/1913475

Hardy, G. H., Littlewood, J. E., & Pólya, G. (1934). *Inequalities* (2nd ed.). Cambridge University Press.

Haughton, J., & Khandker, S. (2009). *Handbook on poverty and inequality*. World Bank Publications. https://openknowledge.worldbank.org/bitstream/handle/10986/11985/9780821376133.pdf

Jann, B. (2016, January). *Estimating Lorenz and concentration curves in Stata* (University of Bern Social Sciences Working Papers No. 15). University of Bern, Department of Social Sciences. https://ideas.repec.org/p/bss/wpaper/15.html

Jenkins, S. (2008). *Estimation and interpretation of measures of inequality, poverty, and social welfare using stata* (North American Stata Users' Group Meetings 2006). Stata Users Group. http://EconPapers.repec.org/RePEc:boc:asug06:16

Kovacevic, M., & Binder, D. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach. *Journal of Official Statistics*, *13*(1), 41–58. http://www.jos.nu/Articles/abstract.asp?article=13141

Krämer, W. (1998). Measurement of inequality. In A. Ullah & D. E. A. Giles (Eds.), *Handbook of applied economic statistics* (1st ed., pp. 39–62). Marcel Dekker.

Langel, M., & Tillé, Y. (2012). Inference by linearization for zenga's new inequality index: A comparison with the gini index. *Metrika*, *75*(8), 1093–1110. https://doi.org/10.1007/s00184-011-0369-1

Lerman, R., & Yitzhaki, S. (1989). Improving the accuracy of estimates of gini coefficients. *Journal of Econometrics*, *42*(1), 43–47. http://EconPapers.repec.org/RePEc:eee:econom:v:42:y:1989:i:1:p:43-47

Lima, L. C. F. (2013, September). *The Persistent Inequality in the Great Brazilian Cities: The Case of Brasília* (MPRA Papers No. 50938). University of Brasília. https://mpra.ub.uni-muenchen.de/50938/

Lumley, T. (2004). Analysis of complex survey samples [R package verson 2.2]. *Journal of Statistical Software*, *9*(1), 1–19.

Lumley, T. (2010). *Complex surveys: A guide to analysis using r: A guide to analysis using r*. John Wiley; Sons.

Lumley, T. (2020). Survey: Analysis of complex survey samples [R package version 4.0].

Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of majorization and its applications* (2nd ed.). Springer.

Medeiros, M. (2006). The Rich and the Poor: The Construction of an Affluence Line from the Poverty Line. *Social Indicators Research*, *78*(1), 1–18. https://doi.org/10.1007/s11205-005-7156-1

Morduch, J. (1998). Poverty, economic growth, and average exit time. *Economics Letters*, *59*(3), 385–390. https://doi.org/10.1016/S0165-1765(98)00070-6

Mosler, K. (1994). Majorization in economic disparity measures [Special Issue Honoring Ingram Olkin]. *Linear Algebra and its Applications*, *199*, 91–114. https://doi.org/https://doi.org/10.1016/0024-3795(94)90343-3

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality. *Journal of the European Survey Research Association*, *3*(3), 167–195. http://ojs.ub.uni-konstanz.de/srm/article/view/369

Peichl, A., Schaefer, T., & Scheicher, C. (2010). Measuring Richness and Poverty: A micro data application to Europe and Germany. *Review of Income and Wealth*, *56*(3), 597–619. https://doi.org/https://doi.org/10.1111/j.1475-4991.2010.00404.x

Ravallion, M. (2016). *The economics of poverty: History, measurement and policy* (1st) [ISBN 10:0190212764, 13:9780190212766]. Oxford University Press.

Rohde, N. (2016). J-divergence measurements of economic inequality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(3), 847–870. https://doi.org/10.1111/rssa.12153

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software*, *79*(10), 1–38. https://doi.org/10.18637/jss.v079.i10

Tillé, Y., & Matei, A. (2021). *Sampling: Survey sampling* [R package version 2.9]. https://CRAN.R-project.org/package=sampling

Watts, H. W. (1968). *An economic definition of poverty* (Discussion Papers No. 5). Institute For Research on Poverty. https://www.irp.wisc.edu/publications/dps/pdfs/dp568.pdf

Wolter, K. M. (1985). *Introduction to variance estimation*. Springer-Verlag.

Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica e Applicazioni*, *1*(4), 3–27.