Poverty and Inequality with Complex Survey Data

Guilherme Jacob, Anthony Damico, and Djalma Pessoa

2016-12-14

Contents

Chapter 1

Introduction

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

For now, you have to install the development version of **bookdown** from Github:

```
devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need to install XeLaTeX.

The library convey aims at estimating measures of poverty and income concentration. There are already at least two libraries covering this subject: vardpoor and Laeken. The main difference between the library convey and these two is that the convey strongly hinges on the survey library.

1.1 Installation

`convey` is free and open-source software that runs inside the R environment for statistical computing.

- the latest released version from CRAN with

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

- the latest development version from github with

  ```
  devtools::install_github("djalmapessoa/convey")
  ```

[This may present how to install R, RStudio and required packages. Providing brief information about `survey` and `MonetDBLite` may also be recommended.]

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1.1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 3.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

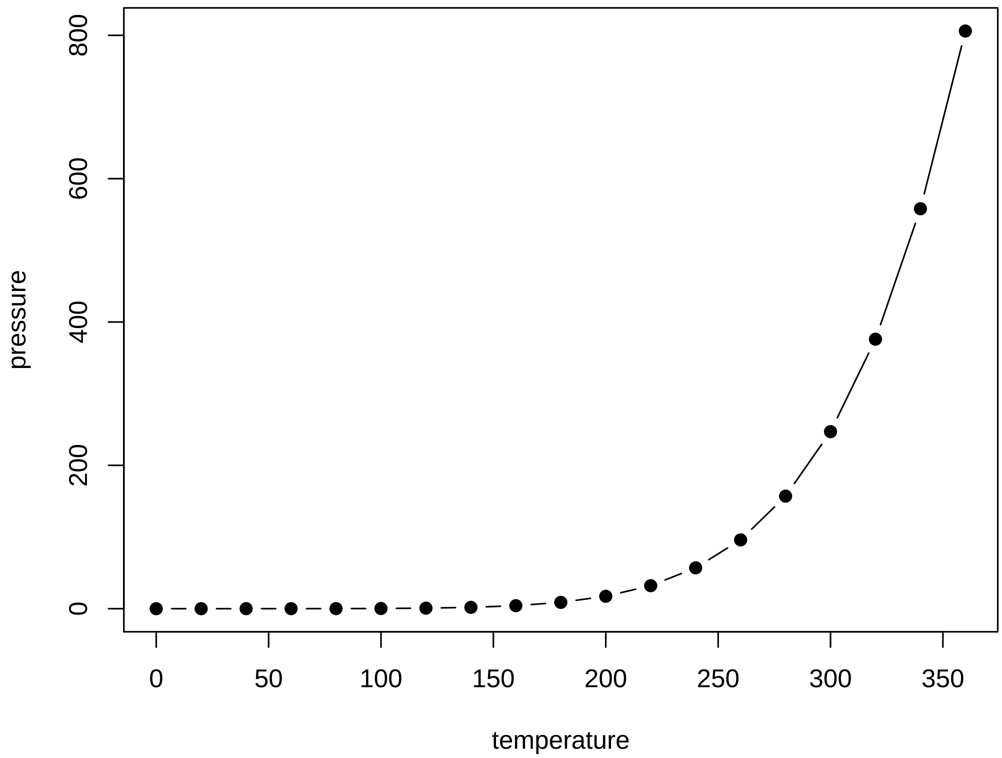

Figure 1.1: Here is a nice figure!

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2016) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

1.2 Complex surveys and statistical inference

In this book we estimate measures of poverty and income concentration in a populatiion, generaly of households or people, based on data colected from a complex survey sample from the population, involving

1- different units selection probabilities;

2- clustering of units;

3- stratification of clusters, and

4- reweighting to compensate missing values and other adjustments.

Items 1 and 4 imply that we should use different units weights to avoid biases when performing statistical analysis. Also, when estimating variances, we should consider, not only the design weights but all listed design characteristics 1-4.

In order to take into account the sample design characteristics it should be used a specialized software like the R library **survey**, adopted in this book.

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

1.3 Linearization

Some measures of poverty and income concentration are defined by non-differentiable functions so that it is not possible to use Taylor linearization to estimate their variances. An alternative is to use **Influence functions** as described in (Deville, 1999) and (Osier, 2009). The library convey 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 income median;
- 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 depend on quantiles and the Gini is defined as a function of ranks. This could be done using the approach proposed by Deville (1999) based upon influence functions.

1.4 Influence function

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: $\hat{F}(x) = T(\hat{M}) = \frac{\int 1(y \leq x) d\hat{M}}{\int d\hat{M}} = \frac{\sum_s w_i 1(y_i \leq x)}{\sum_s w_i}$

1.5 The variance estimator

The variance of the estimator $T(\hat{M})$ can approximated by:

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

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

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

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

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

1.6 Influence functions - Examples

- Total:

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

- Ratio of two totals:

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

1.7 Linearization by influence function - Examples

- 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;

1.8 Structure of the library

In the library convey, there are some basic functions that produces the linearized variables of some estimates that often enter in the definition of measures of concentration and poverty. For example the `quantile` which is linearized by the function `svyiqalpha`. Other example is the function `svyisq` that linearizes the total below a quantile of the variable.

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 rules of composition valid for Gateaux derivatives also hold for Influence Functions.

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

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

In the library convey this rule is implemented by the function `contrastinf` which uses the 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 `arprt`:

$$rmpg = \frac{arpt - medpoor}{arpt}$$

where `medpoor` is the median of incomes less than `arpt`.

Suppose we know how to linearize `arpt` and `medpoor`, then by applying the function `contrastinf` with

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

we linearize the `rmpg`.

1.8.1 Examples of use of the library convey

In the following examples we will use the data set `eusilc` contained in the libraries `vardpoor` and `Laeken`.

```
library(vardpoor)
data(eusilc)
```

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

```
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` for 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 0.017202243
Carinthia         Carinthia 0.1308627 0.010610622
Lower Austria Lower Austria 0.1384362 0.006517660
Salzburg           Salzburg 0.1378734 0.011579280
Styria               Styria 0.1437464 0.007452360
Tyrol                 Tyrol 0.1530819 0.009880430
Upper Austria Upper Austria 0.1088977 0.005928336
Vienna               Vienna 0.1723468 0.007682826
Vorarlberg       Vorarlberg 0.1653731 0.013754670
```

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

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

```
          qsr     SE
eqIncome 3.97 0.0426
```

```
# for domains
svyby(~eqIncome, by = ~db040, design = des_eusilc,
  FUN = svyqsr, alpha= .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)
```

and then use the function `svyarpr`:

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

```
            arpr     SE
eqIncome 0.14444 0.0026
```

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

```
                      db040  eqIncome se.eqIncome
Burgenland       Burgenland 0.1953984 0.015169170
Carinthia         Carinthia 0.1308627 0.011524873
Lower Austria Lower Austria 0.1384362 0.006574264
Salzburg           Salzburg 0.1378734 0.012206087
Styria               Styria 0.1437464 0.007362737
Tyrol                 Tyrol 0.1530819 0.009876583
Upper Austria Upper Austria 0.1088977 0.006449967
Vienna               Vienna 0.1723468 0.007140846
Vorarlberg       Vorarlberg 0.1653731 0.012192981
```

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

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

```
          gini     SE
py010n 0.64606 0.0034
```

djalma, where do these references go on this page? (Berger and Skinner, 2003) and (Osier, 2009) and (Deville, 1999)

Chapter 2

Poverty Indices

[I think this is a good start. I don't think that gender pay gap, quantiles and totals are measures of poverty. Consider another chapter on other wellbeing measures.] this is a test ## At Risk of Poverty Ratio and Threshold (svyarpr, svyarpt)

here are the references

(Osier, 2009) and (Deville, 1999)

2.1 The Gender Pay Gap (svygpg)

here are the references

(Osier, 2009) and (Deville, 1999)

2.2 Quintile Share Ratio (svyqsr)

here are the references

(Osier, 2009) and (Deville, 1999)

2.3 Relative Median Income Ratio (svyrmir)

here are the references

(Osier, 2009) and (Deville, 1999)

2.4 Relative Median Poverty Gap (svyrmpg)

here are the references

(Osier, 2009) and (Deville, 1999)

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

here are the references

(Osier, 2009) and (Deville, 1999)

2.6 Foster-Greer-Thorbecke class (svyfgt)

here are the references

(Foster et al., 1984) and (Berger and Skinner, 2003)

(Foster et al., 1984) proposed a family of indicators to measure poverty.

The 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.

When $\gamma = 0$, p can be interpreted as the poverty headcount ratio, and for $\gamma \geq 1$, the weight of the income shortfall of the poor to a power γ, (Foster and all, 1984).

The poverty measure FGT is implemented in the library convey by the function `svyfgt`. 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 $\theta = .6 * MED$ the measure is equal to the indicator `arpr` computed by the function `svyarpr`.

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 (FGT1) ($\gamma = 1$) 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 FGT0 with the poverty threshold fixed at $0.6 * MED$ we fix the argument type_thresh="relq" and use the default values for `percent` and `order`:

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

```
            fgt0     SE
eqIncome 0.14444 0.0028
```

that matches the estimate obtained by

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

```
            arpr     SE
eqIncome 0.14444 0.0028
```

To estimate the poverty gap(FGT1) 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
```

Chapter 3

Inequality Measurement

3.1 Lorenz Curve (svylorenz)

Though not an inequality measure in itself, the Lorenz curve is a classic instrument of distribution analysis. Basically, it is a function that associates a cumulative share of the population and the share of the total income it owns. In mathematical terms,

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

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).

3.2 Measures derived from the Lorenz Curve

3.2.1 Gini index (svygini)

The Gini index 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:

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

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.

3.2.2 Amato index (svyamato)

The Amato index is also based on the Lorenz curve, but instead of focusing on the area of the curve, it focuses on its length. (Arnold, 2012) proposes a formula not directly based in the Lorenz curve, which (Barabesi et al., 2016) uses to present the following estimator:

$$\widehat{A} = \sum_{i \in S} w_i \left[\frac{1}{\widehat{N}^2} + \frac{y_i^2}{\widehat{Y}^2} \right]^{\frac{1}{2}},$$

which also generates the linearized formula for SE estimation.

The minimum value A assumes is $\sqrt{2}$ and the maximum is 2. In order to get a measure in the interval $[0, 1]$, the standardized Amato index $\widetilde{A}$ can be defined as:

$$\widetilde{A} = \frac{A - \sqrt{2}}{2 - \sqrt{2}} \ .$$

3.2.3 Zenga Index and Curve (svyzenga, svyzengacurve)

The Zenga index and its curve were proposed in (Zenga, 2007). As (Polisicchio and Porro, 2011) noticed, this curve derives directly from the Lorenz curve, and can be defined as:

$$Z(p) = 1 - \frac{L(p)}{p} \cdot \frac{1-p}{1-L(p)}.$$

In the `convey` library, an experimental estimator based on the Lorenz curve is used:

$$\widehat{Z(p)} = \frac{p\widehat{Y} - \widehat{\widehat{Y}}(p)}{p\big[\widehat{Y} - \widehat{\widehat{Y}}(p)\big]}.$$

In turn, the Zenga index derives from this curve and is defined as:

$$Z = \int_0^1 Z(p)dp.$$

However, its estimators were proposed by (Langel, 2012) and (Barabesi et al., 2016). In this library, the latter is used and is defined as:

$$\widehat{Z} = 1 - \sum_{i \in S} w_i \left[\frac{(\widehat{N} - \widehat{H}_{y_i})(\widehat{Y} - \widehat{K}_{y_i})}{\widehat{N} \cdot \widehat{H}_{y_i} \cdot \widehat{K}_{y_i}} \right]$$

where $\widehat{N}$ is the population total, $\widehat{Y}$ is the total income, $\widehat{H}_{y_i}$ is the sum of incomes below or equal to y_i and $\widehat{N}_{y_i}$ is the sum of incomes greater or equal to y_i.

3.3 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 $g(f(y))\colon = -\log f(y)$. Therefore, the expected information or, put simply, the *entropy* is

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

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 the intuition 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 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.

3.3.1 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 top incomes upon the index increases. In some cases, this measure takes special forms, such as mean log deviation and the aforementioned Theil index.

In order to estimate it, (Biewen and Jenkins, 2003) proposed the following:

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

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

This class also has several desirable properties, such as additive decomposition. The additive decomposition allows to compare the effects of inequality within and between population groups on the population inequality. Put simply, an additive decomposable index allows for:

$$I_{Total} = I_{Between} + I_{Within}.$$

3.3.2 Rényi Divergence (svyrenyi)

Another measure used in areas like ecology, statistics and information theory is Rényi divergence measure. Using the formula defined in (Langel, 2012), the estimator can be defined as:

$$\widehat{R}_\alpha = \begin{cases} \frac{1}{\alpha - 1} \log \left[\widehat{N}^{\alpha - 1} \sum_{i \in S} w_i \cdot \left(\frac{y_k}{\widehat{Y}} \right) \right], & \text{if } \alpha \neq 1, \\ \sum_{i \in S} \frac{w_i y_i}{\widehat{Y}} \log \frac{\widehat{N} y_i}{\widehat{Y}}, & \text{if } \alpha = 1, \end{cases}$$

where α is a parameter with a similar economic interpretation to that of the GE_α index.

3.3.3 J-Divergence and Decomposition (svyjdiv, svyjdivdec)

Proposed by (Rohde, 2016), the J-divergence measure can be seen as the sum of GE_0 and GE_1, satisfying axioms that, individually, those two indices do not. Using U_γ and T_γ functions defined in **??**, the estimator can be defined as:

$$\widehat{J} = \frac{1}{\widehat{N}} \sum_{i \in S} w_i \left(\frac{y_i - \widehat{\mu}}{\widehat{\mu}} \right) \log \left(\frac{y_i}{\widehat{\mu}} \right)$$

$$\therefore \widehat{J} = \frac{\widehat{T}_1}{\widehat{U}_1} - \frac{\widehat{T}_0}{\widehat{U}_0}$$

Since it is a sum of two additive decomposable measures, J itself is decomposable.

3.3.4 Atkinson index (svyatk)

Although the original formula was proposed in (Atkinson, 1970), the estimator used here comes from (Biewen and Jenkins, 2003):

$$\widehat{A}_\epsilon = \begin{cases} 1 - \widehat{U}_0^{-\epsilon/(1-\epsilon)} \widehat{U}_1^{-1} \widehat{U}_{1-\epsilon}^{1/(1-\epsilon)}, & \text{if } \epsilon \in \mathbb{R}_+ \setminus \{1\} \\ 1 - \widehat{U}_0 \widehat{U}_0^{-1} exp(\widehat{T}_0 \widehat{U}_0^{-1}), & \text{if } \epsilon \to 1 \end{cases}$$

The ϵ is an inequality aversion parameter: as it approaches infinity, more weight is given to incomes in bottom of the distribution.

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:

```
. 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
---------------------------------------------------------------------------
```

```
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:

```
. 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
---------------------------------------------------------------------------
```

```
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
```

3.3.5 Replicating Barabesi et al. (2016)

Chapter 4

Multidimensional Indices

4.1 Alkire-Foster Class and Decomposition (svyafc, svyafcdec)

In November 2015, Christopher Jindra presented at the Oxford Poverty and Human Development Initiative on the Alkire-Foster multidimensional poverty measure. His presentation can be viewed here. 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 stata-style webuse library
library(webuse)

# load the same microdata set used by Jindra in his presentation
webuse("nlsw88")

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

# create a `collgrad` column
nlsw88$collgrad <-
    factor(
        as.numeric( nlsw88$collgrad ) ,
        label = c( 'not college grad' , 'college grad' ) ,
        ordered = TRUE
      )

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

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

Replicate PDF page 9

```r
page_nine <-
  svyafc(
    ~ wage + collgrad + hours ,
    design = des_nlsw88 ,
    cutoffs = list( 4, 'college grad' , 26 ) ,
    k = 1/3 , g = 0 ,
    na.rm = TRUE
  )

# M0 and seM0
print( page_nine )
```

```
##      alkire-foster     SE
## [1,]       0.36991 0.0053
```

```r
# H seH and A seA
print( attr( page_nine , "extra" ) )
```

```
##        coef          SE
## H 0.8082070 0.008316807
## A 0.4576895 0.004573443
```

Replicate PDF page 10

```r
page_ten <- NULL

# loop through every poverty cutoff `k`
for( ks in seq( 0.1 , 1 , .1 ) ){

    this_ks <-
        svyafc(
            ~ wage + collgrad + hours ,
            design = des_nlsw88 ,
            cutoffs = list( 4 , 'college grad' , 26 ) ,
            k = ks ,
            g = 0 ,
            na.rm = TRUE
           )

    page_ten <-
        rbind(
            page_ten ,
            data.frame(
                k = ks ,
                M0 = coef( this_ks ) ,
                seM0 = SE( this_ks ) ,
                H = attr( this_ks , "extra" )[ 1 , 1 ] ,
                seH = attr( this_ks , "extra" )[ 1 , 2 ] ,
                A = attr( this_ks , "extra" )[ 2 , 1 ] ,
                seA = attr( this_ks , "extra" )[ 2 , 2 ]
          )
        )

}
```

Table 4.1: Here is a nice table!

k	MO	seMO	H	seH	A	seA
0.1	0.3699078	0.0053059	0.8082070	0.0083168	0.4576895	0.0045734
0.2	0.3699078	0.0053059	0.8082070	0.0083168	0.4576895	0.0045734
0.3	0.3699078	0.0053059	0.8082070	0.0083168	0.4576895	0.0045734
0.4	0.1865894	0.0068123	0.2582516	0.0092455	0.7225101	0.0051745
0.5	0.1865894	0.0068123	0.2582516	0.0092455	0.7225101	0.0051745
0.6	0.1865894	0.0068123	0.2582516	0.0092455	0.7225101	0.0051745
0.7	0.0432649	0.0042978	0.0432649	0.0042978	1.0000000	0.0000000
0.8	0.0432649	0.0042978	0.0432649	0.0042978	1.0000000	0.0000000
0.9	0.0432649	0.0042978	0.0432649	0.0042978	1.0000000	0.0000000
1.0	0.0432649	0.0042978	0.0432649	0.0042978	1.0000000	0.0000000

```r
knitr::kable(
  page_ten , caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

still need to replicate PDF page 13

https://github.com/DjalmaPessoa/convey/issues/168

then keep going replicating this

https://github.com/DjalmaPessoa/convey/issues/154

(Alkire and Foster, 2011) and (Sabina Alkire and Ballon, 2015) and (Pacifico and Poge, 2016)

4.2 Bourguignon (1999) inequality class (svybmi)

(Bourguignon, 1999) and (Ana Lugo, 2007)

Bibliography

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, 95(7-8):476–487.

Ana Lugo, M. (2007). *Comparing Multidimensional Indices of Inequality: methods and application*, pages 213–236.

Arnold, B. C. (2012). On the amato inequality index. *Statistics and Probability Letters*, 82(8):1504–1506.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263.

Barabesi, L., Diana, G., and Perri, P. F. (2016). Linearization of inequality indices in the design-based framework. *Statistics*, 50(5):1161–1172.

Berger, Y. G. and 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.

Biewen, M. and Jenkins, S. (2003). Estimation of generalized entropy and atkinson inequality indices from complex survey data. Discussion Papers of DIW Berlin 345, DIW Berlin, German Institute for Economic Research.

Bourguignon, F. (1999). Comment to 'multidimensioned approaches to welfare analysis' by maasoumi, e. In Silber, J., editor, *Handbook of income inequality measurement*, chapter 15, pages 477–484. Kluwer Academic, London.

Cowell, F. A., Flachaire, E., and Bandyopadhyay, S. (2009). Goodness-of-fit: An economic approach. Economics Series Working Papers 444, University of Oxford, Department of Economics.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25(2):193–203.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, 52(3):761–766.

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.

Kovacevic, M. and Binder, D. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach. *Journal of Official Statistics*, 13(1):41–58.

Langel, M. (2012). *Measuring inequality in finite population sampling*. PhD thesis.

Lerman, R. and Yitzhaki, S. (1989). Improving the accuracy of estimates of gini coefficients. *Journal of Econometrics*, 42(1):43–47.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality. *Journal of the European Survey Research Association*, 3(3):167–195.

Pacifico, D. and Poge, F. (2016). Mpi: Stata module to compute the alkire-foster multidimensional poverty measures and their decomposition by deprivation indicators and population sub-groups.

Polisicchio, M. and Porro, F. (2011). A comparison between lorenz l(p) curve and zenga i(p) curve. *Statistica Applicata*, 21(3-4):289–301.

Rohde, N. (2016). J-divergence measurements of economic inequality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(3):847–870.

Sabina Alkire, James Foster, S. S. M. E. S. J. M. R. and Ballon, P. (2015). *Multidimensional Poverty Measurement and Analysis*. Oxford University Press. ISBN 9780199689491.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.3.

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.