Measuring Inequality in Sampling Surveys

Sebastian Wójcik

Statistics Poland

Presentation plan

- Introduction
- Inequality measures
- R package
- Example

Introduction

- One of the core activities of official statistics is conducting the sample surveys.
- An integral part of sample surveys is a sampling error. It has a strong impact on reliability of provided estimates in terms of precision.
- Information about estimated totals, means etc. is incomplete without their precision or confidence interval.
- Sampling error is strictly connected with sampling design, inclusion probabilities and sampling weights.
- Methods used for processing the sample surveys must enable a data weighting.
- Data weighting is an important element in calculating the estimates' precision.

Introduction

- Concept of inequality is in a domain of interest of economists, sociologists, statisticians etc.
- Income Inequality, Lifetime Inequality, Inequality of Wealth and Inequality of Opportunity are particular concepts of inequality in economics.
- Scientists developed many inequality measures for variables on ratio scale as well as for ordinal scale.

Introduction

- Among various R packages for data analysis, not all of them implement methods suitable for weighted data.
- In a case of inequality analysis, there is lack of packages that offer a wide spectrum of inequality measures as well as data weighting and provide precision of a given inequality measure.
- It lead us to idea of creating relevant package built up on new content as well as on using and extending functions in available packages (ineq, sampling)
- After testing and enhancing, the R package will be available on GitHub account of Statistics Poland https://github.com/statisticspoland

Inequality measures

The R package contains several measures of inequality:

- available in ineq package: Atkinson, Gini, Kolm, Ricci-Schutz, Theil, Entropy and Coefficient of Variation
- new measures: Hoover index, Jenkins index, Cowell and Flachaire index, Palma ratio, 20:20 ratio, Leti index, Allison and Foster index.

All methods were implemented with the possibility to weight the data.

Inequality measures

Assume that:

x_i - value of the i-th element

w_i - weight of the i-th element

μ - arithmetic mean

μ_w- weighted arithmetic mean

$$\mu_w = \frac{\sum w_i x_i}{\sum w_i}$$

Inequality measures

Gini coefficient

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu} \qquad G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|w_i w_j}{2(\sum_{i=1}^{n} w_i)^2\mu_w}$$

Hoover Index

$$H = \frac{1}{2}\frac{\sum_{i=1}^{n}|x_i - \mu|}{\sum_{i=1}^{n} x_i} \qquad H = \frac{1}{2}\frac{\sum_{i=1}^{n} w_i|x_i - \mu_w|}{\sum_{i=1}^{n} w_i x_i}$$

Inequality measures

Theil index

Theil T

$$T_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{\mu}\ln\frac{x_i}{\mu}$$

$$T_1 = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i\frac{x_i}{\mu_w}\ln\frac{x_i}{\mu_w}$$

Theil L

$$T_0 = \frac{1}{n}\sum_{i=1}^{n}\ln\frac{\mu}{x_i}$$

$$T_0 = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i\ln\frac{\mu_w}{x_i}$$

Inequality measures

Atkinson's coefficient

Determined for a ϵ parameter chosen by the researcher

$$A = \begin{cases} 1 - \frac{1}{\mu}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}} & for\ \epsilon \neq 1 \\ 1 - \frac{1}{\mu}\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} & for\ \epsilon = 1 \end{cases}$$

$$A = \begin{cases} 1 - \frac{1}{\mu_w}\left(\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i x_i^{1-\epsilon}\right)^{\frac{1}{1-\epsilon}} & for\ \epsilon \neq 1 \\ 1 - \frac{1}{\mu_w}\left(\prod_{i=1}^{n} {x_i}^{w_i}\right)^{\frac{1}{\sum_{i=1}^{n} w_i}} & for\ \epsilon = 1 \end{cases}$$

Inequality measures

Let us denote:

- $c = (c_1, c_2, \cdots, c_m)$ - ordinal scale of values ($c_i < c_j$, whenever $i <$ j) e.g. Likert scale,
- n_i - frequency of c_i. In a case of weighted data $n_j = \sum_{k:\, x_k = c_j} w_k$
- $F_N^i = \frac{\sum_{j=1}^{i} n_j}{n}$.

Leti index

$$L = 2 \sum_{i=1}^{m-1} F_N^i \left[1 - F_N^i\right]$$

Inequality measures

Allison and Foster Index

k - the median category derived from the relations $c_k = med(X)$

Mean value below median

$$\mu_X^L(c) = 2\left(\sum_{i=1}^{k-1} c_i \left(F_N^i - F_N^{i-1}\right) + c_k\left(0{,}5 - F_N^{k-1}\right)\right)$$

Mean value above median

$$\mu_X^U(c) = 2\left(\sum_{i=k+1}^{n} c_i \left(F_N^i - F_N^{i-1}\right) + c_k\left(F_N^k - 0{,}5\right)\right)$$

Allison and Foster index

$$I_X^{AF}(c) = \mu_X^U(c) - \mu_X^L(c)$$

R package

The R package was developed in the following steps:

- Deriving inequality measures formulas for weighted data.
- Coding new inequality measures
- Extending inequality measures from ineq package to cover weighted data
- Adding function *ineq.weighted* to deliver all inequality measures for a given data set
- Adding function *ineq.weighted.boot* to extend the output of *ineq.weighted* by bootstrap

R package

The package contains the following inequality measures:

- Hoover(X,W=rep(1,length(X)),
- Gini(X,W=rep(1,length(X)),
- Theil_L(X,W=rep(1,length(X)),
- Theil_T(X,W=rep(1,length(X)),
- Atkinson(X,W=rep(1,length(X),Atkinson.e=1),
- Kolm(X,W=rep(1,length(X),parameter=1),
- Entropy(X,W=rep(1,length(X),parameter=0.5),
- CoefVar(X,W=rep(1,length(X)),
- RicciSchutz(X,W=rep(1,length(X)),
- Leti(X,W=rep(1,length(X)),
- AF(X,W=rep(1,length(X)),
- Prop20_20(X,W=rep(1,length(X)),
- Palma(X,W=rep(1,length(X)),
- Jenkins(X,W=rep(1,length(X),Jenkins.alpha=0.8) (covers also Cowell and Flachaire index)

X is a data vector, W is a vector of weights.

R package

```
ineq.weighted(
X,
W=rep(1,length(X)),
Atkinson.e=1,
Jenkins.alfa=0.8,
Entropy.e=0.5,
Kolm.p=1)
```

Output
The function generates weighted mean and sum of X, and all inequality measures.

R package

```
ineq.weighted.boot(
X,W=rep(1,length(X)),
Atkinson.e=1,Jenkins.alfa=0.8,Entropy.e=0.5,Kolm.p=1,
keepSamples=F,
keepMeasures=F,
B=1000,
conf.alpha=0.05,
calib.boot=F,
Xs=rep(1,length(X)),
total=sum(W),
calib.method='truncated')
```

B – numer of bootstrap samples.
keepSamples – if TRUE, it returns bootstrap samples of data (Xb) and weights (Wb)
keepMeasures – if TRUE, it returns values of all inequality measures for each bootstrap sample
calib.boot – if FALSE, then naive bootstrap is performed, calibrated bootstrap elsewhere
calib.method – weights' calibration method for function *calib* (sampling)
conf.alpha – significance level for confidence interval.

R package

```
ineq.weighted.boot(
X,W=rep(1,length(X)),
Atkinson.e=1,Jenkins.alfa=0.8,Entropy.e=0.5,Kolm.p=1,
keepSamples=F,
keepMeasures=F,
B=1000,
conf.alpha=0.05,
calib.boot=F,
Xs=rep(1,length(X)),
total=sum(W),
calib.method='truncated')
```

Output

For weighted mean and weighted total of X as well as for each inequality measure, this functions returns outputs from *ineq.weighted* and bootstrap outcomes: expected value, bias [in %], standard deviation, coefficient of variation, lower and upper bound of confidence interval.

Example

Ilocos {ineq} - Income metadata from surveys conducted by the Philippines' National Statistics Office. Weights included.

We shall compare income inequality among men and women (head of a household).

```
library(ineq)
data(Ilocos);attach(Ilocos)
Female <- ineq.weighted.boot(income[sex=='female'],
                             AP.weight[sex=='female'],
                             keepMeasures = T,B = 1000)
Male <- ineq.weighted.boot(income[sex=='male'],
                           AP.weight[sex=='male'],
                           keepMeasures = T,B = 1000)
```

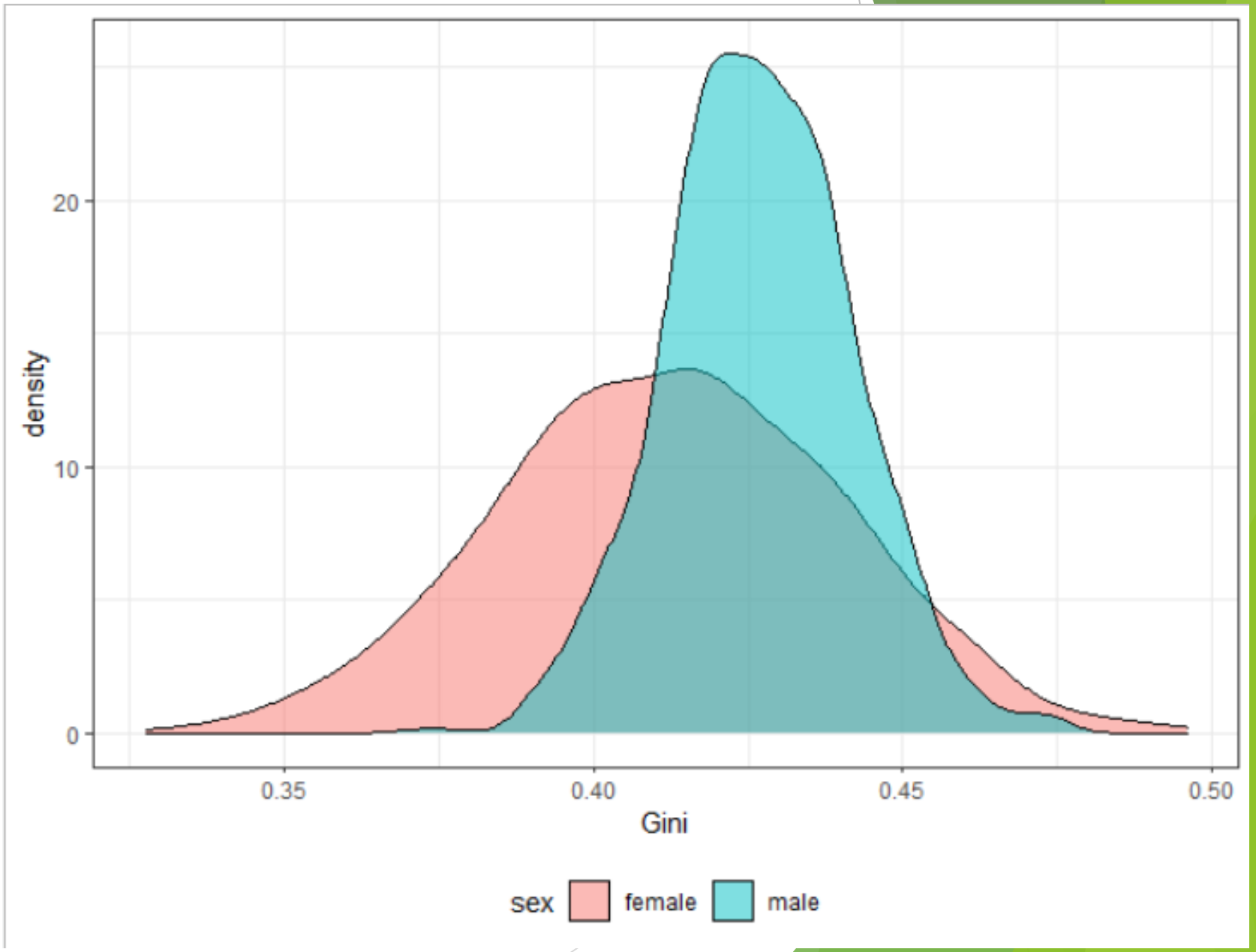
Example

Visualize Gini coefficient distributions with ggplot2.

```
library(ggplot2)
df=data.frame(sex=c(rep(c('female','male'),each=1000)),
              Gini=c(Female[[2]][,'Gini'],Male[[2]][,'Gini']))
df %>% ggplot()+
  geom_density(aes(x=Gini,fill=sex),alpha=0.5)+
  theme_bw()+
  theme(legend.position = 'bottom')
```

Example

Visualize Gini coefficient distributions with ggplot2.



Example

Testing normality of distribution with Jarque-Bera test from normtest package.

In both cases p-value is greater than 0.05.

Hence, we can use statistical tests which are based on normal distribution assumption.

```
library('normtest')
library('dplyr')
Female[[2]][,'Gini'] %>% jb.norm.test()
```

```
##
##  Jarque-Bera test for normality
##
## data:  .
## JB = 2.5024, p-value = 0.257
```

```
Male[[2]][,'Gini'] %>% jb.norm.test()
```

```
##
##  Jarque-Bera test for normality
##
## data:  .
## JB = 1.9303, p-value = 0.3775
```

Example

Testing difference in means with *t.test*. P-value is lower than 0.05. Mean value of Gini coefficient for women amounted to 0.4129 and it is significantly lower than for men which amounted to 0.42678

```
t.test(Female[[2]][,'Gini'],Male[[2]][,'Gini'])
```

```
##
##  Welch Two Sample t-test
##
## data:  Female[[2]][, "Gini"] and Male[[2]][, "Gini"]
## t = -14.106, df = 1559.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01571447 -0.01187777
## sample estimates:
## mean of x mean of y
## 0.4129748 0.4267709
```

Thank you for attention

Sebastian Wójcik
s.wojcik@stat.gov.pl

Statistics Poland