Variance estimation with the R package gustave: the experience of STATEC

Claude LAMBORAY, Dmitri LEBRUN, Guillaume OSIER

uRos2023, Bucarest

12-14 December 2023

STATEC

1. Variance estimation methodology for EU-SILC
2. Implementation with R package Gustave
3. Preliminary results (EU-SILC 2022)

1

Variance estimation methodology for EU-SILC

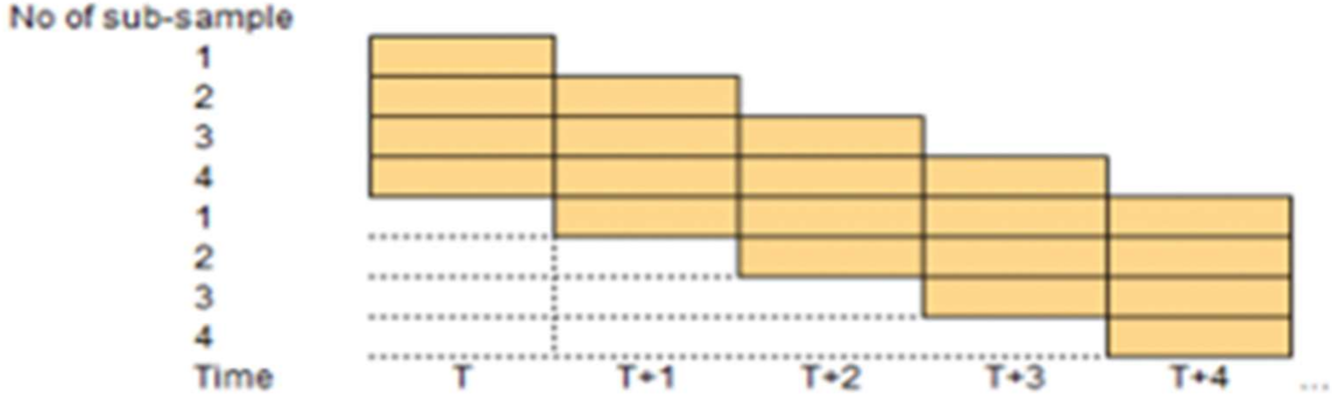
STATEC

EU-SILC

- EU-SILC : *European Statistics on Income and Living Conditions*
- Reference source at EU level for comparable microdata on income and living conditions across countries
- Used to produce key policy indicators on income poverty, inequality and social exclusion
- Strong legal basis at EU level (IESS Framework Regulation on Social Statistics) setting minimum precision targets
- Complex design features:
  - Rotating panel structure
  - Unequal selection probabilities
  - Unit non-response and attrition
  - Calibration to external data sources
  - Both cross-sectional and longitudinal indicators
  - Both linear and non-linear indicators

Illustration of the 4-year rotational design

The 4-year rotational design refers to a sample selection based on a number of 4 sub-samples or replications, each of them similar in size and design and representative of the whole population. From one year to the next, three out of four replications are retained, while the first selected one (the "oldest" one) is dropped and replaced by a new replication.



According to this design, the cross-sectional component of EU-SILC operation in year T is composed of four sub-samples each of them drawn in different years: T-3, T-2, T-1 and T.

The need to calculate standard error estimates (IESS Regulation)

- Variance estimates are key indicators for assessing the quality of survey data
- According to the IESS Regulation at EU level, a minimum level of precision must be achieved for EU-SILC indicators

ANNEX II

Precision requirements

1. Precision requirements for all data sets are expressed in standard errors and are defined as continuous functions of the actual estimates and of the size of the statistical population in a country or in a NUTS 2 region.
2. The estimated standard error of a particular estimate $\widehat{\mathrm{SE}}(\hat{p})$ shall not be bigger than the following amount:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{f(N)}}$$

3. The function f(N) shall have the form of $f(N)=a\sqrt{N}+b$

- Therefore, standard errors must be calculated using an approach that is both theoretically justified sound and easy to implement using standard statistical software

Variance decomposition (3-phase sampling)

$$\widehat{V}(\widehat{Y}) = \widehat{V}\left(\sum_{i \in s_3} \frac{y_i}{\pi_i p_i r_i}\right)$$

$$= \sum_{\substack{i \in s_3 \\ j \in s_3}} \frac{A_{ij}}{\pi_{ij}\pi_i\pi_j} \times \frac{y_i}{p_i r_i} \times \frac{y_j}{p_j r_j} + \sum_{i \in s_3} \frac{1-\pi_i}{\pi_i^2} \times \left(\frac{1}{p_i r_i} - \frac{1}{p_i^2 r_i^2}\right) \times y_i^2$$

Initial Sampling (Phase 1)

$$+ \sum_{\substack{i \in s_3 \\ j \in s_3}} \frac{y_i^2}{\pi_i^2} \times \frac{1-p_i}{p_i^2} \times \frac{1}{r_i}$$

Unit non-response (Phase 2)

$$+ \sum_{\substack{i \in s_3 \\ j \in s_3}} \frac{y_i^2}{\pi_i^2 p_i^2} \times \frac{1-r_i}{r_i^2}$$

Attrition (Phase 3)

Additional indicators

- Standard error (absolute): $\hat{\sigma}(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}$

- Standard error (relative): $\widehat{CV}(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}/\hat{Y}$

- Margin of error at 95% confidence level (absolute): $\widehat{AME}(\hat{Y}) = 1.96\sqrt{\hat{V}(\hat{Y})}$

- Margin of error at 95% confidence level (relative): $\widehat{RME}(\hat{Y}) = 1.96\sqrt{\hat{V}(\hat{Y})}/\hat{Y}$

- Design effect: ratio of the variance under the actual sampling plan to that under simple random sampling of same size

$$Deff = \frac{\hat{V}(\hat{Y})}{\hat{V}_{SRS}(\hat{Y}_{SRS})} = \frac{\hat{V}(\hat{Y})}{\frac{1}{r} \times \hat{N}^2 \times \left(1 - \frac{r-1}{\hat{N}-1}\right) \times \left(\sum_{i \in s_3} \omega_i (y_i - \bar{y}_w)^2 / \sum_{i \in s_3} \omega_i\right)}$$

2

Application in gustave

STATEC

R package gustave

- Gustave: a User-oriented Statistical Toolkit for Analytical Variance Estimation
- R package developed at INSEE https://github.com/InseeFr/gustave
- We decided to work with this package because:
  - Analytical variance methods are typically used at STATEC.
  - The package is designed to separate the technical (methodological) part from the user part.
  - The package is very flexible, allowing to take into account the characteristics of each survey.
  - The package facilitates some degree of standardization.
  - The package is used within an official statistics environment.

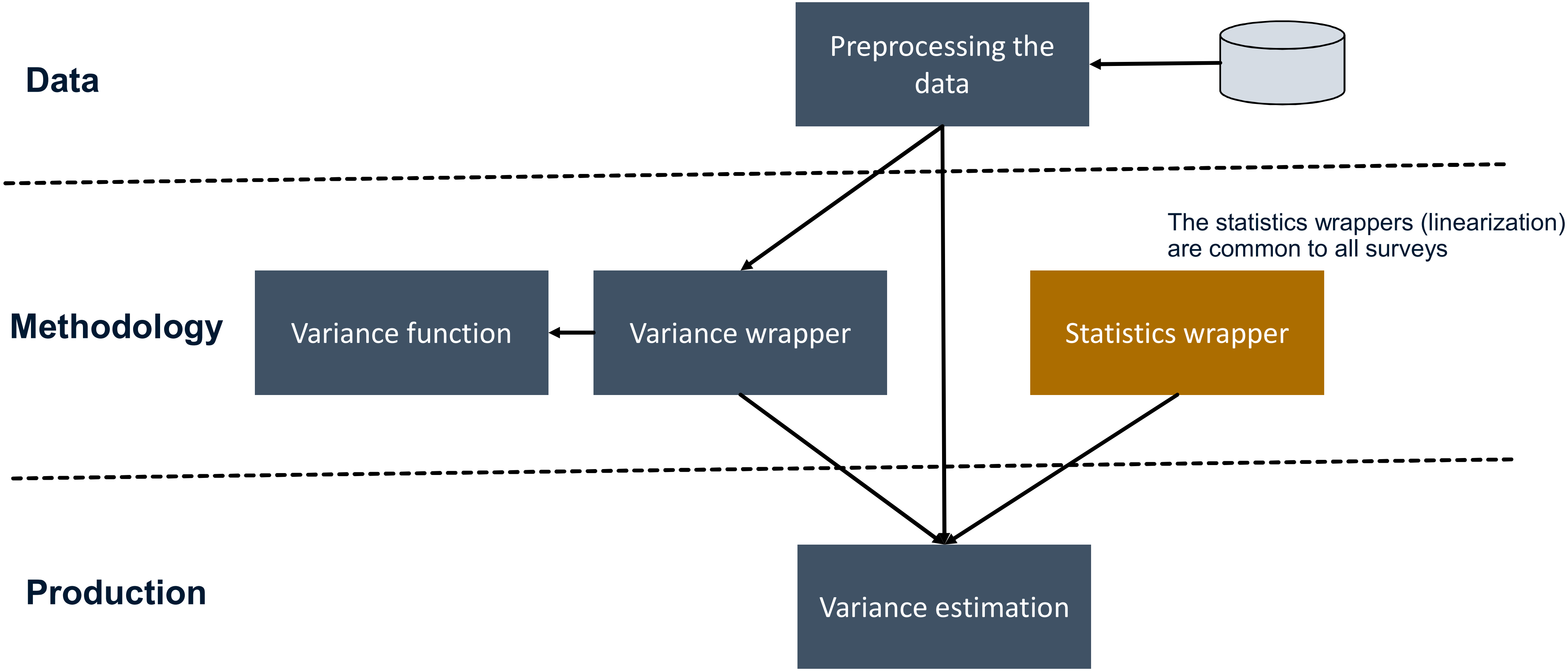
Getting started with *qvar*

- gustave contains a ready-to-be-used function *qvar* that allows to obtain a variance wrapper for some common situations:
  - Stratified Simple Random Sampling
  - Non-response correction
  - Calibration

- The function *qvar* creates a **variance wrapper** that can be called by the end user

```
precision_Crime_cal <- qvar(
  #data set
  data=crimedb2,
  #disseimantion weight
  dissemination_dummy = "resp",
  #dissemination weight
  dissemination_weight = "r2_w",
  #identification variable
  id = "ID",
  #sampling weight
  sampling_weight = "w_sample",
  #weight after non-response correction
  nrc_weight = "w_nrc",
  #response dummy
  response_dummy = "resp",
  #weight after calibration
  calibration_weight = "r2_w",
  #calibration variables
  calibration_var = unlist(VecteurGustave),
  #define a variance wrapper
  define = TRUE
)
```

Application of gustave at STATEC



Variance wrapper

The variance wrapper requires:

- A variance function;
- The technical data (weights, id, calibration variables, etc.);
- The unit identifier;
- The final weight used for point estimation;
- The name of the variable that identifies the units.

```
Calcul_variance_SILC <- gustave:::define_variance_wrapper(
  variance_function = VarSILC,
  technical_data = technical_data_SILC,
  reference_id = technical_data_SILC$data$id_hfile,
  reference_weight = technical_data_SILC$data$weight_final,
  default_id = 'id_hfile'
)
```

STATEC

Variance function

- The variance estimation methodology

is specified in the variance function.

- The variance function returns

a variance for a TOTAL.

- We can use supporting functions

(var_pois, var_DT, …) available in gustave to

estimate the variances

- In addition to the variance, we decided

to calculate other outputs:

    - Design effect (eds);
    - Sampling variance (var1); Non-response variance (var2); Attrition variance (var3).

```
VarSILC <- function(y, data){
  ....
  ....
  variance[['p1']] <- varDT(y = y_pp / (data$Proba_rep * data$Attrition_proba),
                            pik = 1 / data$household_design_weight,
                            strata = data$strate)

  variance[['p2']] <- var_pois(y = y_pp,
                               pik = data$Attrition_proba * data$Proba_rep,
                               w = data$household_design_weight)

  ....
  ....
  return(list(var = Reduce(`+`, variance),
              eds = sqrt(Reduce(`+`, variance) / VarSRS),
              var1 = Reduce(`+`, variance) - var_nr - var_attr,
              var2 = var_nr,
              var3 = var_attr))
}
```

Statistics wrapper

- A statistic wrapper returns the point estimator and the corresponding linearized variable.
- The display (output of the variance wrapper) is also managed in the statistics wrapper.
- We constructed additional statistics wrappers for:
  - Quantile
  - ARPR
  - RMPG
  - Gini
  - S80/S20 quintile share ratio

```
RATIO_STATEC <- gustave:::define_statistic_wrapper(
  statistic_function = function(num, denom, weight){
    na <- is.na(num) | is.na(denom)
    est_num <- sum(num * weight, na.rm = TRUE)
    est_denom <- sum(denom * weight, na.rm = TRUE)
    point <- est_num / est_denom
    lin <- (num - point * denom ) / est_denom
    list(point = point,
         lin = lin,
         n = sum(!na),
         est_num = est_num,
         est_denom = est_denom)
  },
  arg_type = list(data = c("num", "denom"),
                  weight = "weight"),

  display_function = STATEC_standard_display_function
)
```

Variance estimation

- The variance wrapper can be called by specifying the **data set** and the **variable of interest** included in the corresponding **statistics wrapper**

```
> Calcul_variance_SILC(Data_SILC_Gustave, MEAN_STATEC(HY020))
                  call    n      est variance      std       cv    lower    upper      moe relative_moe
1 MEAN_STATEC(y = HY020) 3911 78976.99  3075589 1753.736 2.220565 75539.73 82414.25 3437.259        4.352
  effet_de_sondage var_tirage_share var_nr_share var_attr_share
1            2.083           13.833       68.938         17.229
```

- The variance wrapper conveniently handles **domain estimation.**

```
> Calcul_variance_SILC(Data_SILC_Gustave, MEAN_STATEC(HY020), by=hs031)
                             call by    n      est variance      std        cv    lower    upper       moe
1 MEAN_STATEC(y = HY020, by = hs031)  1   62 54188.44 39559210 6289.611 11.606923 41861.03 66515.86 12327.412
2 MEAN_STATEC(y = HY020, by = hs031)  2  101 54297.85 44503374 6671.085 12.286095 41222.76 67372.93 13075.086
3 MEAN_STATEC(y = HY020, by = hs031)  3 3709 80726.70  3419978 1849.318  2.290838 77102.10 84351.30  3624.597
  relative_moe effet_de_sondage var_tirage_share var_nr_share var_attr_share
1       22.749            1.551           16.374       79.744          3.883
2       24.080            1.645           15.381       71.658         12.961
3        4.490            0.456           13.790       68.458         17.753
```

3

Preliminary results (EU-SILC 2022)

STATEC

Preliminary results (SILC 2022)

	Value	Standard error (points of unit value)	CV (%)	Absolute margin of error (points of unit value)	Relative margin of error (%)	Deff
60% of median income (*at-risk-of-poverty line*) – EUR/month	2265	42	1.9	83	3.7	1.59
Share of individuals below the poverty line (*at-risk-of-poverty rate*) – %	17.3	1.1	6.4	2.2	12.7	1.58
Relative difference between the poverty line and the median income of the poor (*relative median at-risk-of-poverty gap*) – %	18.2	1.9	10.5	3.7	20.4	1.62
Mean income of the upper income quintile to the median income of the lower income quintile (*income quintile share ratio*)	4.5	0.5	10.3	0.9	20.2	1.67
Gini coefficient (%)	29.1	1.4	4.9	2.8	9.5	4.5

Conclusion and way forward

- Integrated approach that yields variance estimates taking into account the EU-SILC complex design features

- The R package Gustave provides both the theoretical foundations and the flexibility in implementation thanks to the statistics and variance wrappers; however some pre-processing remains necessary

- Work in progress
  - Further validation of results
  - Effect of calibration weighting on variance
  - Application to other surveys (e.g. Labour Force Survey, other business and household surveys etc.)
  - Inclusion of other indicators (longitudinal indicators and indicators of net changes)