Survey Sampling Methodology in Latvia

Mārtiņš Liberts

Central Statistical Bureau of Latvia

8–12 December 2013

Content

Content

Central Statistical Bureau of Latvia

- Established on the 1st September 1919
- Incorporation in statistical system of Soviet Union 1945
- Independence regained in 1991
- 549 employees at the beginning of 2013
- The main provider of the official statistics in Latvia
- Survey methodology is used since 90-ties.

Mathematical Support Division

- Survey methodology is applied in centralised manner
- Mathematical Support Division is responsible for:
  - Sampling design and sampling
  - Weighting of survey data
  - Imputation (only for social surveys)
  - Precision estimation (sampling errors)

Mathematical Support Division

- Survey methodology is applied in centralised manner
- Mathematical Support Division is responsible for:
  - Sampling design and sampling
  - Weighting of survey data
  - Imputation (only for social surveys)
  - Precision estimation (sampling errors)
  - Time series analyses:
    - Seasonal adjustment
    - Short term forecasting
  - Training of the CSB staff

Content

Software Used for Survey Methodology

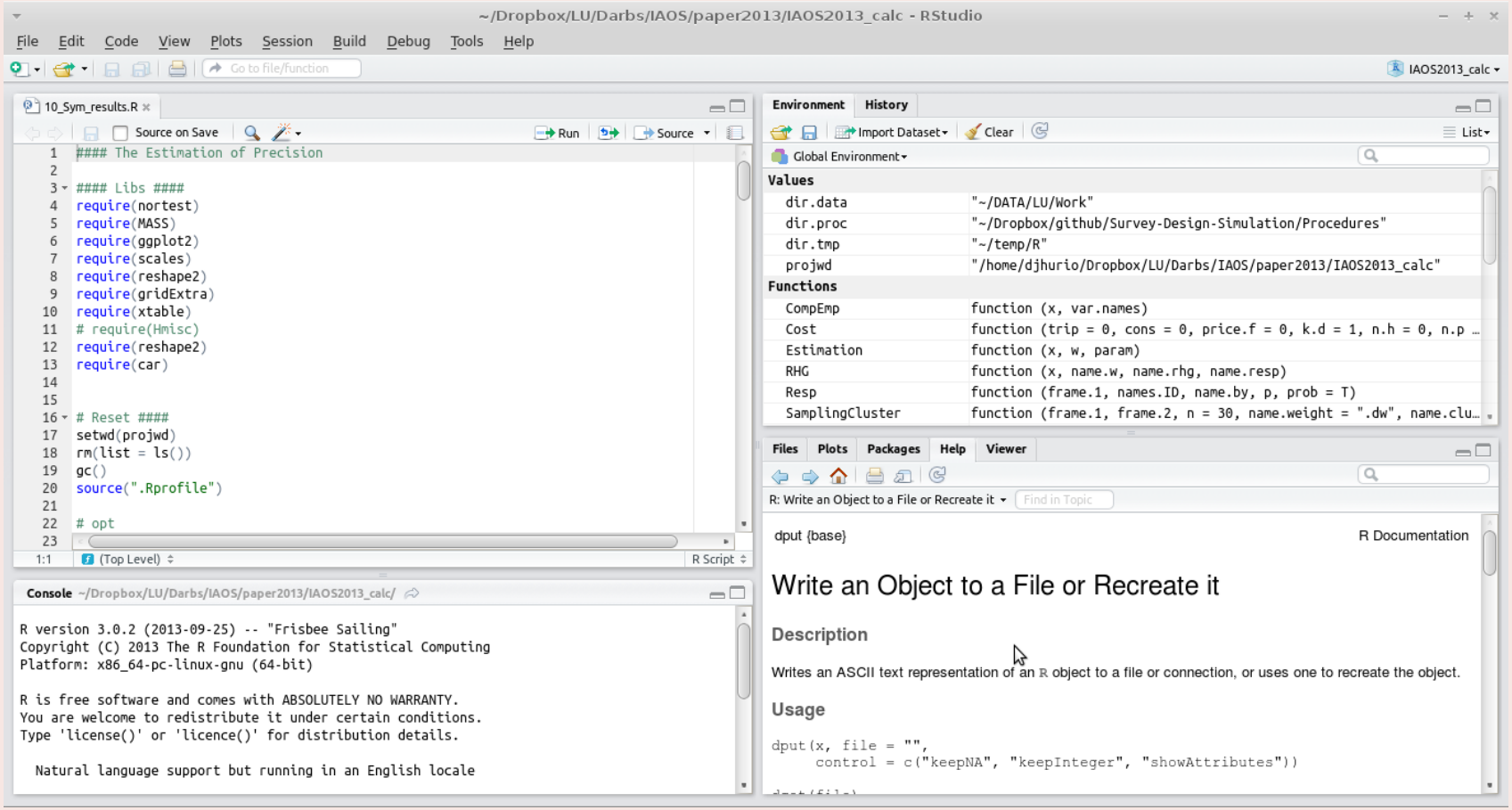
- SPSS
  - Sampling
  - Weighting
  - Imputation
- R (`http://www.r-project.org/`)
  RStudio (`http://www.rstudio.com/`)
  - Sampling
  - Weighting, calibration
  - Imputations
  - Sampling error estimation
- Demetra+, JDemetra+
  - Seasonal Adjustment

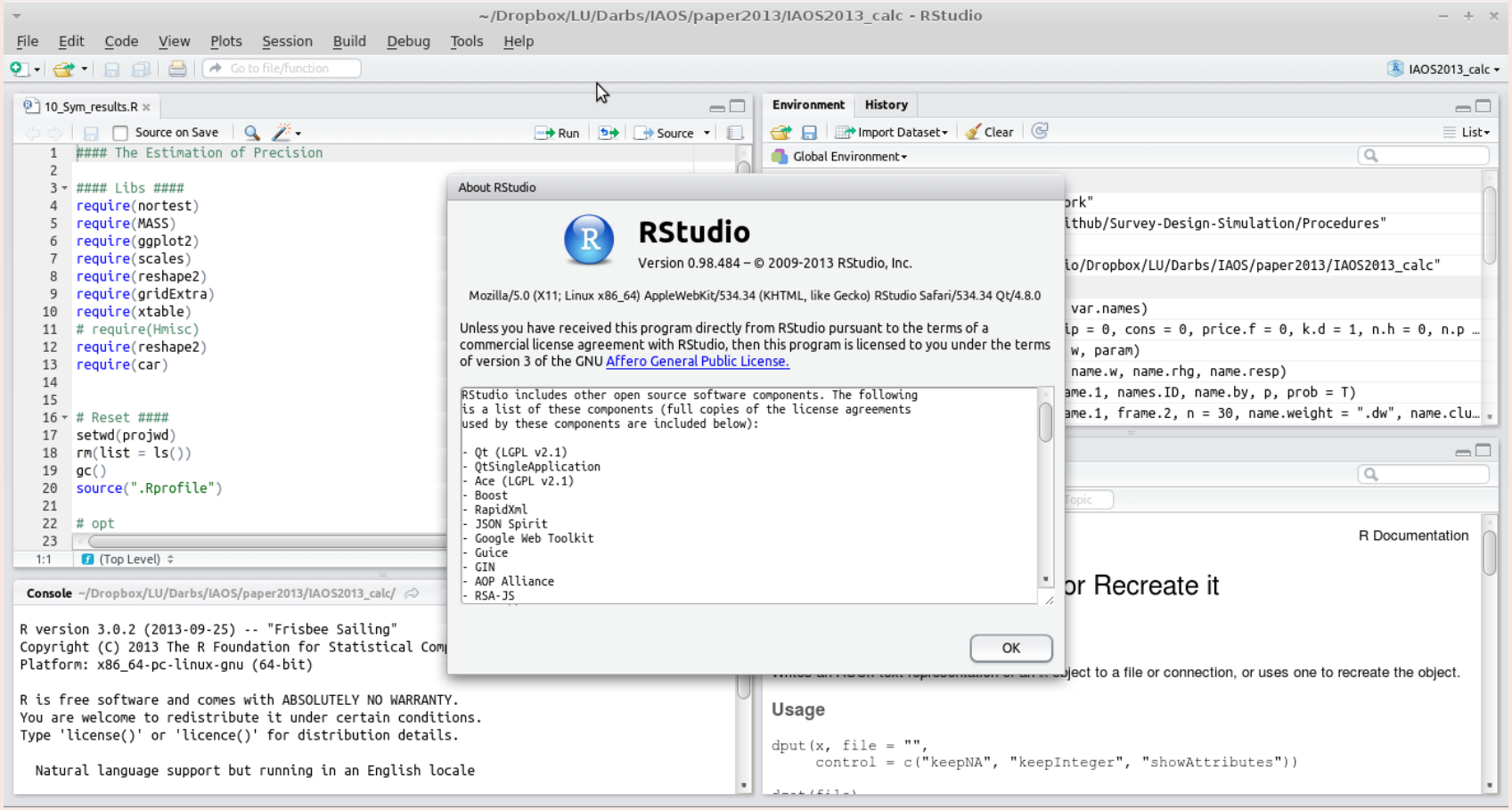
R examples

- Sampling
  - function `dom_optimal_allocation`
  - Jānis Jukāms, Central Statistical Bureau of Latvia
- Calibrations
  - package `sampling` function `calib`
  - Yves Tillé and Alina Matei (2012). sampling: Survey Sampling. R package version 2.5. `http://CRAN.R-project.org/package=sampling`
- Sampling error estimation
  - package `vardpoor` function `vardom`
  - Juris Breidaks, Mārtiņš Liberts (2013) Central Statistical Bureau of Latvia

R examples

- Sampling error estimation
  - package `vardpoor` function `vardom`
  - Juris Breidaks, Mārtiņš Liberts (2013) Central Statistical Bureau of Latvia


~/Dropbox/LU/Darbs/IAOS/paper2013/IAOS2013_calc - RStudio
File Edit Code View Plots Session Build Debug Tools Help
IAOS2013_calc
10_Sym_results.R
Source on Save
Run
Source
1 #### The Estimation of Precision
2
3 #### Libs ####
4 require(nortest)
5 require(MASS)
6 require(ggplot2)
7 require(scales)
8 require(reshape2)
9 require(gridExtra)
10 require(xtable)
11 # require(Hmisc)
12 require(reshape2)
13 require(car)
14
15
16 # Reset ####
17 setwd(projwd)
18 rm(list = ls())
19 gc()
20 source(".Rprofile")
21
22 # opt
23
1:1 (Top Level) R Script
Console ~/Dropbox/LU/Darbs/IAOS/paper2013/IAOS2013_calc/
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
Environment History
Import Dataset
Clear
List
Global Environment
Values
dir.data "~/DATA/LU/Work"
dir.proc "~/Dropbox/github/Survey-Design-Simulation/Procedures"
dir.tmp "~/temp/R"
projwd "/home/djhurio/Dropbox/LU/Darbs/IAOS/paper2013/IAOS2013_calc"
Functions
CompEmp function (x, var.names)
Cost function (trip = 0, cons = 0, price.f = 0, k.d = 1, n.h = 0, n.p …
Estimation function (x, w, param)
RHG function (x, name.w, name.rhg, name.resp)
Resp function (frame.1, names.ID, name.by, p, prob = T)
SamplingCluster function (frame.1, frame.2, n = 30, name.weight = ".dw", name.clu…
Files Plots Packages Help Viewer
R: Write an Object to a File or Recreate it
Find in Topic
dput {base} R Documentation
Write an Object to a File or Recreate it
Description
Writes an ASCII text representation of an R object to a file or connection, or uses one to recreate the object.
Usage
dput(x, file = "",
control = c("keepNA", "keepInteger", "showAttributes"))


~/Dropbox/LU/Darbs/IAOS/paper2013/IAOS2013_calc - RStudio
File Edit Code View Plots Session Build Debug Tools Help
10_Sym_results.R
Source on Save
Run
Source
The Estimation of Precision
Libs
require(nortest)
require(MASS)
require(ggplot2)
require(scales)
require(reshape2)
require(gridExtra)
require(xtable)
# require(Hmisc)
require(reshape2)
require(car)
# Reset ####
setwd(projwd)
rm(list = ls())
gc()
source(".Rprofile")
# opt
(Top Level)
Console ~/Dropbox/LU/Darbs/IAOS/paper2013/IAOS2013_calc/
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
Environment
History
Import Dataset
Clear
List
Global Environment
About RStudio
RStudio
Version 0.98.484 – © 2009-2013 RStudio, Inc.
Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/534.34 (KHTML, like Gecko) RStudio Safari/534.34 Qt/4.8.0
Unless you have received this program directly from RStudio pursuant to the terms of a commercial license agreement with RStudio, then this program is licensed to you under the terms of version 3 of the GNU Affero General Public License.
RStudio includes other open source software components. The following is a list of these components (full copies of the license agreements used by these components are included below):
- Qt (LGPL v2.1)
- QtSingleApplication
- Ace (LGPL v2.1)
- Boost
- RapidXml
- JSON Spirit
- Google Web Toolkit
- Guice
- GIN
- AOP Alliance
- RSA-JS
OK
R Documentation
or Recreate it
ject to a file or connection, or uses one to recreate the object.
Usage
dput(x, file = "",
control = c("keepNA", "keepInteger", "showAttributes"))

Content

Sampling Frame in Household Surveys

- Source: Statistical Dwelling Register
  - Population Register
  - Building Register
  - Address Register
  - other
- Output for sampling:
  - List of census counting areas
  - List of dwellings
  - List of persons

Sampling in Household Surveys

- Sampling design depends on mode of survey:
  - Paper assisted **personal** interviews (PAPI)
  - Computer assisted **personal** interviews (CAPI)
  - Computer assisted **telephone** interviews (CATI)
  - Computer assisted **web** interviews (CAWI)
- CAPI is a traditional mode
- The usage of CATI is increasing
- CAPI or CAPI/CATI are the most common modes
- CAWI was used in the last population census (2011)

Sampling in Household Surveys

- There are travelling costs if CAPI is used
- We want to minimize / optimise the travelling costs
- Two-stage sampling design is used:
  - Sampling of (*census counting*) areas is used in the 1st stage
  - Sampling of dwellings or individuals is used in the 2nd stage
- This allows to reduce / control travelling costs

Sampling in Household Surveys

- There are travelling costs if CAPI is used
- We want to minimize / optimise the travelling costs
- Two-stage sampling design is used:
  - Sampling of (*census counting*) areas is used in the 1st stage
  - Sampling of dwellings or individuals is used in the 2nd stage
- This allows to reduce / control travelling costs
- **Attention**: The sampling errors tend to increase because of a clustering effect

Sampling in Household Surveys (other designs)

- Stratified simple random sampling (SSRS)
  - Survey on doctoral degree holders (very small population)
- Two-stage sampling for CAPI and SSRS for CATI
  - European Health Survey (2014)

Sampling Frame in Economic Surveys

- Source: Statistical Business Register
  - State Business Register
  - State Revenue Service (tax office)
  - other
- Output for sampling:
  - List of active enterprises and organisations

Sampling in Economic Surveys

- CAWI (e-questionnaire) is the most common mode
- postal surveys
- call back is used in all cases if necessary

Sampling in Economic Surveys

- CAWI (e-questionnaire) is the most common mode
- postal surveys
- call back is used in all cases if necessary
- **Remark:** There are no travelling costs

Sampling in Economic Surveys

- Stratified simple random sampling is used in most cases
- Stratification variables:
  - Size groups (size measured by turnover, number of employees)
  - Economic activity branch (NACE classification)
  - Type of unit
  - Region
- Optimal sample allocation for each domain (*R procedure*)

Sampling in Economic Surveys (Other Designs)

- Two stage sampling for the Survey on Employees:
  - Units (enterprises or local units) sampled in the first stage
  - Employees sampled in the second stage

Sampling Frame in Agriculture Surveys

- Source: Statistical Farm Register
  - State Land Service
  - Farming Land Register
  - Animal Register
  - other
- Output for sampling:
  - List of active farms

Sampling in Agriculture Surveys

- CAWI (e-questionnaire) is used as the first mode
- CAPI is used as the second mode

Sampling in Agriculture Surveys

- CAWI (e-questionnaire) is used as the first mode
- CAPI is used as the second mode
- **Remark:** There are travelling costs (in CAPI mode)

Sampling in Agriculture Surveys

- Stratified simple random sampling is used in most cases
- Stratification variables:
  - Size groups (size measured by land area, economic size)
  - Specialisation
  - Region

Content

Weighting in Household Surveys

Usual scheme:

- Design weights according to sampling design (computed during sampling)
- Non-response correction by response homogeneity groups
- Trimming of extreme weights (optional)
- Calibration of weights to external information (population counts)
  - Function `calib()` from the `sampling` package (R)

Weighting in Household Surveys (Other Cases)

- Cross-sectional and longitudinal weighting for European Survey on Income and Living Conditions (EU-SILC)

Weighting in Economic and Agriculture Surveys

Usual scheme:

- Design weights according to sampling design (computed during sampling)
- Usually population frame is updated (sampling frame and weighting frame)
- Weighting to population counts:
  - Number raised estimator in each stratum
  - Calibration of weights

Content

Imputation in Household Surveys

- Usually imputation is done for income variables
- Methods:
  - Randomised hot-deck imputation in groups
  - Nearest neighbour (distance function) imputation
  - Regression:
    - Linear regression
    - Interval regression

Imputation in Economic and Agriculture Surveys

- Done by subject matter unit (decentralised)
- Case specific (large enterprises are heterogeneous)
- Imputation using:
  - Administrative records
  - Historical data

Content

Estimation of Sampling Error

- Common procedure for all surveys
- Written as an R procedure – function `vardom` from package `vardpoor`
- Main steps of the procedure:
  - Domain variables are generated for domain estimates

$$y_{i,d} = \begin{cases} y_i & \text{if} \quad i \in D_d \\ 0 & \text{if} \quad i \notin D_d \end{cases}$$

  - Linearised variables are computed for non-linear statistics (ratio of two totals, Gini index, ...)
  - Residuals from the regression model are estimated if weight calibration is applied
  - Ultimate cluster variance estimator (Hansen, Hurwitz and Madow, 1953)

Ultimate cluster variance estimator

- Ultimate cluster variance estimator (Hansen, Hurwitz and Madow, 1953)
- Osier (2012) "The Linearisation Approach Implemented by Eurostat for The First Wave of EU-SILC: What Could Be Done from The Second Wave Onwards?"

$$\hat{V}\left(\hat{\Theta}\right) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(z_{hi\bullet}-\bar{z}_{h\bullet\bullet}\right)^2$$

$$z_{hi\bullet} = \sum_{j=1}^{m_{hi}} \omega_{hij}\cdot z_{hij}$$

$$\bar{z}_{h\bullet\bullet} = \frac{\sum_{i=1}^{n_h} z_{hi\bullet}}{n_h}$$

Input for `vardom`

- `Y` – study variables
- `H` – stratification
- `PSU` – ID of primary sampling units
- `w_final` – final weight
- `Dom` – domain variables
- `N_h` – PSU population size in each stratum
- `Z` – denominator variables (for ratio)
- `X` – calibration variables
- `g` – calibration factor (g-weight)

Output from `vardom`

- `estim` – parameter estimator
- `var` – variance
- `se` – standard error
- `cv` – coefficient of variation
- `CI_lower` – the lower bound if confidence interval
- `CI_upper` – the upper bound if confidence interval
- `deff` – design effect

Content

Organisation of Household Surveys

- Information Technology Division – updates the Statistical Dwelling Register (SDR)
- Mathematical Support Division (MSD) – extracts information from SDR to build the sampling frame (SPSS syntax)
- MSD creates sample file
- Sample file is sent to:
  - Interview Organisation Division
  - Employment Statistics Division (Labour Force Survey)
  - Income Statistics Division (Survey on Income and Living Standards)

Organisation of Economic Surveys

- Business Register Division (BRD) – updates the Statistical Business Register
- Mathematical Support Division (MSD) – receives a file from BRD with active units (enterprises, organisations)
- MSD creates sample file
- Sample file is sent to subject matter divisions

Organisation of Agriculture Surveys

- Information Technology Division (ITD) – updates the Statistical Farm Register
- Mathematical Support Division (MSD) – receives a file from ITD with active farms
- MSD creates sample file
- Sample file is sent to Agriculture Statistics Division

Thank you!