Small Area Estimation with R

Unit 2: Design-based estimators

V. Gómez-Rubio

Department of Mathematics
Universidad de Castilla-La Mancha, Spain

useR! 2016
27th June 2016, Stanford University

Design-based estimation

Definition (Rao, 2003)

In the context of sample surveys, we refer to a domain estimator as "direct" if it is based only on the domain-specific sample data.

(...)

Design based estimators make use of survey weights, and the associated inferences are based on the robability distribution induced by the sampling design with the population values held fix (...).

Survey design and estimation (Särndall et al., 2003)

- The goal of a survey is to get information about unknown **population characteristics** or **parameters**.
- A survey concerns a finite set of **elements** called a **finite population**. Such subpopulations re called **domains of study** or just **domains**.
- A value of one or more **variables of study** is associated with each population element.
- Access to and observation of individual population elements is established through a **sampling frame**, a device that associates the elements of the population with the **sampling units** in the frame.
- From the population, a **sample** is selected. A sample is a **probability sample** if realized by a chance mechanism.
- For each element in the sample the variables of study are **measured** and the values **recorded**.
- The record variable values are used to calculate (**point**) **estimates** of the finite population parameters of interest (total, means, medians, ratios, regression coefficients, etc.)

Example: Labour Force Survey (Särndall et al., 2003)

```
How many persons are currently in the labor force in the
country as a whole and in various regions of the country?
How many are unemployed?
```

- **Population**: All persons in the country with certain exceptions (such as infants, people in institutions)
- **Domains of interest**: age/sex groups of the population, occupation groups in the population, and regions of the country.
- **Variables**: Each person can be described at the time of the survey as
  - Belonging to the force survey or not
  - Employed or not
- **Population characteristics of interest**: Number of persons in the labor force. Number of persons unemployed in the labor force. Proportion of persons unemployed in the labor force.
- **Sample**: Obtained in an efficient manner.
- **Data processing and estimation**

Survey sampling

Once the sampling frame has been established, the units to be included in the sample can be chosen in different ways:

- Simple random sampling (without replacement)
- Sistematic sampling
- Clustered sampling
- Two-stage sampling
- More complex survey designs

Some problems that may occur while sampling:

- Non-response
- Selection bias

R packages for direct estimation and sampling

`sampling`

- Functions for drawing sampling and calibration
- Implements a wealth of sampling schemes
- Horvitz-Thomson and calibration estimators

`survey`

- Provides methods to analyse data obtained from complex surveys
- Summary statistics and graphics
- Methods available include generalised linear models, post-stratification, calibration and raking

The MSU284 Population

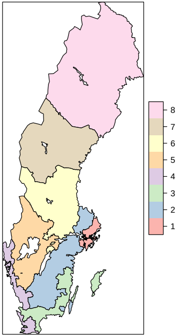
The MSU284 Population (Särndal et al., 2003) describes the 284 municipalities of Sweden. It is included in package `sampling`.

- LABEL. Identifier.
- P85. Population in 1985
- RMT85. Revenues from the 1985 municipal taxation
- ME84. Number of Municipal Employees in 1984
- REG. Geographic region indicator (8 regions)
- CL. *Cluster* indicator (50 clusters)

```
> library(sampling)
> data(MU284)
> MU284 <- MU284[order(MU284$REG), ]
> MU284$LABEL <- 1:284
> summary(MU284)
```

Regions in Sweden

- Municipalities in Sweden can be grouped into 8 regions
- We will treat the municipalities as the *units*
- To estimate the regional mean we will sample from the municipalities
- **Warning!!** It has not been possible to merge the map to the the MU284 data set, but it does not matter for the purpose of this example



Survey sampling with R

Simple Random Sampling Without Replacement

- Sample is made of 32 municipalities ($\sim$11% sample)
- Equal probabilities for all municipalities

```
> #Select a few areas (Estimation of the national revenues)
> N <- 284 #Total number of municipalities
> n <- 32    #~1% Sample size
> nreg <- length(unique(MU284$REG))
> #Simple random sampling without replacement
> set.seed(1)
> smp <- srswor(n, N)
> dsmp <- MU284[smp == 1, ]
> table(dsmp$REG)

1 2 3 4 5 6 7 8
2 5 6 3 7 3 2 4
```

Survey sampling with R

Stratified SRS Without Replacement

- Sample is made of 32 municipalities ($\sim$11% sample)
- 4 municipalities sampled per region
- Equal probabilities for all municipalities **within** strata (i.e., region)

```
> #Multi-stage random sampling
> set.seed(1)
> smpcl <- mstage(MU284, stage = list("cluster", "cluster"),
+    varnames = list("REG", "LABEL"),
+    size = list(8, rep(4, 8)), method = c("srswor", "srswor") )
> dsmpcl <- MU284[smpcl[[2]]$LABEL, ]
> table(dsmpcl$REG)

1 2 3 4 5 6 7 8
4 4 4 4 4 4 4 4
```

Survey sampling with R

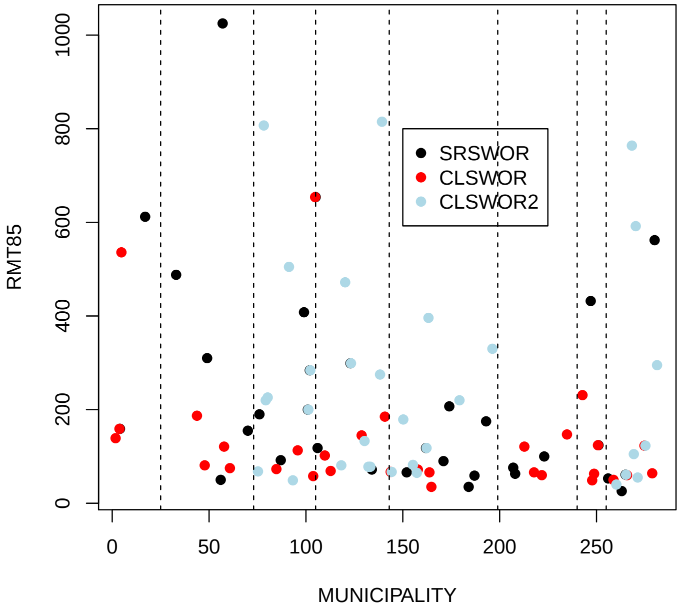
Stratified SRS Without Replacement (Two-Stage Sampling)

- Sample is made of 32 municipalities (∼11% sample)
- 8 municipalities sampled per region
- Equal probabilities for all municipalities **within** strata
- Some regions do not contribute to the survey sample

```
> #Multi-stage random sampling WITH MISSING AREAS
> set.seed(1)
> smpcl2 <- mstage(MU284, stage = list("cluster", "cluster"),
+    varnames = list("REG", "LABEL"),
+    size = list(4, rep(8, 8)), method = c("srswor", "srswor") )
> dsmpcl2 <- MU284[smpcl2[[2]]$LABEL, ]
> table(dsmpcl2$REG)

3 4 5 8
8 8 8 8
```

Survey sampling with R



Direct Estimation

Horvitz-Thomson estimator

- Direct estimators rely on the survey sample to provide small area estimates
- Not appropriate if there are out-of-sample areas

Horvitz-Thomson estimator:

$$\hat{Y}_{direct} = \sum_{i \in s} \frac{1}{\pi_i} y_i \qquad \hat{\bar{Y}}_{direct} = \sum_{i \in s} \frac{\frac{1}{\pi_i} y_i}{\sum_{i \in s} \frac{1}{\pi_i}}$$

For SRS without replacement: $\pi_i = \frac{n}{N}$

The following code computes some summary results that we will use later to assess the quality of the estimates:

```
> library(survey)
> RMT85 <- mean(MU284$RMT85)
> RMT85REG <- as.numeric(by(MU284$RMT85, MU284$REG, mean))
>
```

Direct Estimation

Estimation using SRSWR:

```
> svy <- svydesign(~ 1, data = dsmp, fpc = rep(284, n))
> dest <- svymean(~ RMT85, svy, deff = TRUE)
> #destvar<-svyvar(~RMT85, svy)
```

Estimation using two-stage sampling:

```
> fpc <- lreg[dsmpcl$REG]
> svycl <- svydesign(id = ~ 1, strata = ~ REG, data = dsmpcl, fpc = fpc)
> destcl <- svymean(~ RMT85, svycl, deff = TRUE)
> #destclvar<-svyvar(~RMT85, svycl)
```

Estimation using two-stage sampling from 4 regions:

```
> fpc2 <- lreg[dsmpcl2$REG]
> svycl2 <- svydesign(id = ~ 1, strata = ~ REG, data = dsmpcl2,
+     fpc = fpc2)
> destcl2 <- svymean(~ RMT85, svycl2, deff = TRUE)
> #destcl2var<-svyvar(~RMT85, svycl2)
>
```

Direct Estimation of Domains

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$\hat{\bar{Y}}_{direct} = \sum_{i \in s} \frac{\frac{1}{\pi_i} y_i}{\sum_{i \in s} \frac{1}{\pi_i}}$$

```
> #Estimation of domains
> destdom <- svyby( ~ RMT85,  ~ REG, svy, svymean)
> #destdomvar<-svyby(~RMT85, ~REG, svy, svyvar)
> destdom
```

```
  REG     RMT85         se
1   1 385.50000 153.281044
2   2 405.60000 146.880066
3   3 304.66667  71.773416
4   4 163.00000  54.140880
5   5 107.14286  21.270148
6   6  79.66667   8.468488
7   7 278.00000 104.217576
8   8 175.50000 106.961194
```

Direct Estimation of Domains

A domain refers to a subpopulation of the area of interest

In the example, we may estimate the revenues for each region

$$\hat{\bar{Y}}_{direct} = \sum_{i \in s} \frac{\frac{1}{\pi_i} y_i}{\sum_{i \in s} \frac{1}{\pi_i}}$$

```
> destdomcl <- svyby(~ RMT85, ~ REG, svycl, svymean)
> #destdomclvar<-svyby(~RMT85, ~REG, svycl, svyvar)
> destdomcl

  REG   RMT85          se
1   1 1774.25 1373.886834
2   2  116.00   24.677363
3   3  224.50  134.359553
4   4  125.25   23.905952
5   5   60.00    8.129166
6   6   98.50   20.146264
7   7  116.75   35.467689
8   8   74.25   15.333341
```

Direct Estimation of Domains

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$\hat{\bar{Y}}_{direct} = \sum_{i \in s} \frac{\frac{1}{\pi_i} y_i}{\sum_{i \in s} \frac{1}{\pi_i}}$$

```
> destdomcl2 <- svyby(~ RMT85, ~ REG, svycl2, svymean)
> #destdomcl2var<-svyby(~RMT85, ~REG, svycl2, svyvar)
> destdomcl2

  REG   RMT85       se
3   3 294.875 76.57462
4   4 278.875 81.07735
5   5 182.125 41.06296
8   8 254.375 83.41048
```

Problems of direct estimation

- Direct estimation is only useful if we collect a sample from every domain of interest
- Estimates have usually very wide variances
- What if we hve covariates? Is there any way of improving the estimates?
- What can we say about unsampled domains?

Generalised Regression Estimator

Definition

- Model-assisted estimator
- Relies on survey design and (linear) regression
- It can be expressed as a direct estimator plus some correction term based on additional information (covariates)

$$\hat{Y}_{GREG} = \sum_{j \in s} \frac{1}{\pi_j} y_j + \sum_k \beta_k \left(\sum_{p=1}^{N} x_p - \sum_{j \in s} \frac{1}{\pi_j} x_j \right)$$

$$\hat{Y}_{GREG,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij} + \sum_k \beta_k \left(\sum_{p=1}^{N_i} x_p - \sum_{j \in s_i} \frac{1}{\pi_{ij}} x_{ij} \right)$$

Coefficients β_k are estimated using weigthed linear regression.

GREG Estimation with R

```
> pop.totals = c((Intercept) = N, ME84 = sum(MU284$ME84))
> svygreg<-calibrate(svy, ~ ME84, calfun = "linear",
+    population = pop.totals )
> svymean(~ RMT85, svygreg)

        mean     SE
RMT85 237.58 4.2859

> svygregcl <- calibrate(svycl, ~ ME84, calfun = "linear",
+    population = pop.totals )
> svymean(~ RMT85, svygregcl)

        mean     SE
RMT85 240.03 3.0741

> svygregcl2 <- calibrate(svycl2, ~ ME84, calfun = "linear",
+    population = pop.totals )
> svymean(~ RMT85, svygregcl2)

       mean     SE
RMT85 240.8 3.2212

>
```

Other types of estimators

Post-stratification

- An 'external' source is used to obtain the weights and these are used in the computation of the direct estimator
- Direct standardisation in epidemiology is an example:
  - Population data is available per gender and age group
  - Age/sex mortality/morbidity rates are obtained from the national government, WHO, etc.
  - Expected counts can be computed by combining these two data sources

Other R packages

- The `Social Sciences` Task View (available on CRAN) may provide more information on packages for the collection and analysis of survey data
- `spsurvey`
  This group of functions implements algorithms required for design and analysis of probability surveys such as those utilized by the U.S. Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP).
- `reweight`
  Adjusts the weights of survey respondents so that the marginal distributions of certain variables fit more closely to those from a more precise source (e.g. Census Bureau's data).
- `surveyNG`
  Complex survey samples – database interface, sparse matrices.

Comparing different sampling schemes and estimators

Empiricial Mean Square Error

It is used to assess the quality of Small Area Estimators:

$$AEMSE = \frac{1}{K}\sum_{i=1}^{K}(\hat{\overline{Y}}_i - \overline{Y}_i)^2$$

Design effect

The design effetc is used to compare the variability of the same estimator for a particular sampling scheme $p(s)$. Usually, SRS is taken as the reference:

$$DEFF_{p(s)} = \frac{V_{p(s)}[\hat{\overline{Y}}]}{V_{SRS}[\hat{\overline{Y}}]}$$

Example: Comparing different sampling schemes

The following table shows the results computed with the methods described in this section:

	AEMSE	DEFF
NAT. SRS	224.828378	1.0000000
NAT. CL	157.639519	0.6304145
NAT. CL2	5.995211	0.9543790

Example: Comparing different sampling schemes