EIROPAS SOCIĀLAIS FONDS

EIROPAS SAVIENĪBA

LATVIJAS UNIVERSITATE ANNO 1919

IEGULDĪJUMS TAVĀ NĀKOTNĒ

The Simulation Study of Survey Cost and Precision

Martins Liberts

University of Latvia
Central Statistical Bureau of Latvia

28 August 2012

This work has been supported by the European Social Fund within the project **"Support for Doctoral Studies at University of Latvia - 2"**.

Outline

Outline

Introduction

The target population

Design cost efficiency

Simulation

Conclusions

Motivation

- The idea of the study comes from purely practical necessity
- A balance between precision and cost
- Cost efficiency is a desirable property for sample surveys done in practice.

For example, cluster sampling can be preferable choice regarding cost efficiency because the reduction of cost can dominate the loss in precision.

Tasks of the research

- Define cost efficiency for sampling design
- Propose a tool for measuring or estimating cost efficiency for sampling design

The work is based on Labour Force Survey

Outline

The target population of the Labour Force Survey (LFS)

- All residents permanently living in private households
- Individuals in working age – 15-74 is the main domain of the interest
- Observed on weekly bases (continuously) by the methodology of LFS

The target population

i	w=1	w=2	w=3	w=4	w=5	$\cdots$	w=W
1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$\cdots$	$y_{1,W}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$\cdots$	$y_{2,W}$
3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$\cdots$	$y_{3,W}$
4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$\cdots$	$y_{4,W}$
5	$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$\cdots$	$y_{5,W}$
6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$	$\cdots$	$y_{6,W}$
$\cdots$							
N	$y_{N,1}$	$y_{N,2}$	$y_{N,3}$	$y_{N,4}$	$y_{N,5}$	$\cdots$	$y_{N,W}$

- An assumption of fixed set of individuals during the period of W weeks

Parameter of interest – total

- Weekly total

$$Y_w = \sum_{i=1}^{N} y_{i,w} \tag{1}$$

- Quarterly total

$$Y_q = \frac{1}{13} \sum_{w=j}^{j+12} Y_w = \frac{1}{13} \sum_{w=j}^{j+12} \sum_{i=1}^{N} y_{i,w} \tag{2}$$

- Yearly total

$$Y_y = \frac{1}{4} \sum_{q=k}^{k+3} Y_q = \frac{1}{52} \sum_{w=j}^{j+51} Y_w = \frac{1}{52} \sum_{w=j}^{j+51} \sum_{i=1}^{N} y_{i,w} \tag{3}$$

Parameter of interest – ratio of two totals

- Weekly ratio of two totals

$$R_w = \frac{Y_w}{Z_w} = \frac{\sum_{i=1}^{N} y_{i,w}}{\sum_{i=1}^{N} z_{i,w}} \tag{4}$$

- Quarterly ratio of two totals

$$R_q = \frac{Y_q}{Z_q} = \frac{\sum_{w=j}^{j+12} Y_w}{\sum_{w=j}^{j+12} Z_w} \tag{5}$$

- Yearly ratio of two totals

$$R_y = \frac{Y_y}{Z_y} = \frac{\sum_{q=k}^{k+3} Y_q}{\sum_{q=k}^{k+3} Z_q} = \frac{\sum_{w=j}^{j+51} Y_w}{\sum_{w=j}^{j+51} Z_w} \tag{6}$$

Outline

Precision

- Population parameter θ
- Probability sample s drawn by known sampling design $p(s)$
- θ can be estimated using an estimator $\hat{\theta}_p$
- The variance of $\hat{\theta}_p$ is denoted by $V\left(\hat{\theta}_p\right)$

Cost

- Cost associated to a sample s
- Money, time or other quantity
- Cost function $c(s)$
- Cost of sample s can be computed by the cost function $c_s = c(s)$
- c_s is a random because s is a random sample (for example travelling costs if survey is done by personal interviewing)
- The expectation of c_s under sampling design $p(s)$ is notated as $E(c_s) = C_p$

The balance of precision and cost

- Minimise $V\left(\hat{\theta}_p\right)$ and C_p
- But:
  - $V\left(\hat{\theta}\right)\downarrow\Rightarrow C\left(\hat{\theta}\right)\uparrow$
  - $C\left(\hat{\theta}\right)\downarrow\Rightarrow V\left(\hat{\theta}\right)\uparrow$
- sampling design $p\left(s\right)$ so that C_s and $V\left(\hat{\theta}_p\right)$ would be in "balance"

Design effect

- Measure of design cost efficiency regarding precision and cost
- Assume two sampling designs:
  - Simple random sampling – srs
  - Alternative sampling design – $p(s)$

Design effect

The classical design effect

$$deff\,(p,\theta,n) = \frac{V\left(\hat{\theta}_p \middle| E\,(n_p) = n\right)}{V\left(\hat{\theta}_{srs} \middle| n_{srs} = n\right)} \tag{7}$$

Design effect

The classical design effect

$$deff\left(p,\theta,n\right)=\frac{V\left(\hat{\theta}_p\middle|E\left(n_p\right)=n\right)}{V\left(\hat{\theta}_{srs}\middle|n_{srs}=n\right)} \quad (7)$$

Alternative design effect

$$deff^{\star}\left(p,\theta,\gamma\right)=\frac{V\left(\hat{\theta}_p\middle|C_p=\gamma\right)}{V\left(\hat{\theta}_{srs}\middle|C_{srs}=\gamma\right)} \quad (8)$$

Definitions

Definition

The sampling design $p(s)$ *is more cost efficient then the sampling design* $q(s)$ *for estimation of* θ *with survey budget* γ *if* $deff^{\star}\left(p,\hat{\theta},\gamma\right) < deff^{\star}\left(q,\hat{\theta},\gamma\right)$.

Definition

The sampling design $p(s)$ *is more cost efficient then the sampling design* $q(s)$ *for estimation of* θ *with survey budget* γ *if* $V\left(\hat{\theta}_p, C_p = \gamma\right) < V\left(\hat{\theta}_q, C_q = \gamma\right)$.

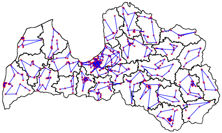
Outline

Sampling designs

- Three sampling designs:
  - SRS of individuals evenly distributed by weeks (only sampled individual takes part in survey)
  - SRS of dwellings evenly distributed by weeks (all individuals from sampled dwelling takes part in survey) = cluster sample of individuals
  - Two stage sampling design used in practice for LFS (Liberts 2010)

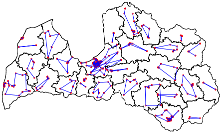
SRS Individuals

Sample size = 1032; Trip = 5048 (km)



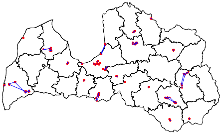
Cluster Sampling of persons

Sample size = 464; Trip = 3242 (km)



Two Stage Sampling of Dwellings

Sample size = 464; Trip = 487 (km)



Simulation

- Artificial population data (created from the Population Register and LFS)
- Cost function

Cost function

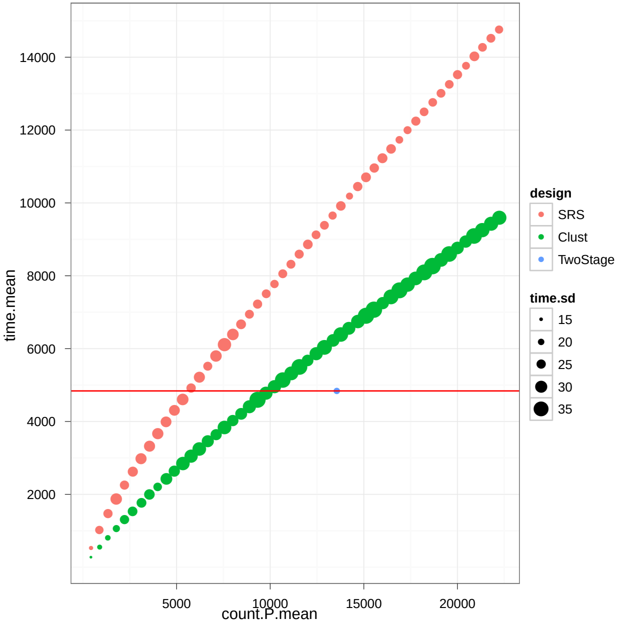
The cost is expressed as time necessary for field interviewers to carry out the survey in the simulation. There are two components:

- Time for travelling $t_1(s) = \frac{\sum_{g=1}^{G} d_g}{\bar{v}}$ where:
  - G is a number of interviewers
  - d_g is a distance done by interviewer g to carry out the surveys
  - $\bar{v}$ – an average travelling speed of interviewer
- Time for interviewing $t_2(s) = m \cdot \bar{t}_H + n \cdot \bar{t}_P$ where:
  - m is number of dwellings taking part in survey
  - n is the number of individuals taking part in survey
  - $\bar{t}_H$ is an average time for a household interview
  - $\bar{t}_P$ is an average time for a personal interview

$$c(s) = t_1(s) + t_2(s) = \frac{\sum_{g=1}^{G} d_g}{\bar{v}} + m \cdot \bar{t}_H + n \cdot \bar{t}_P \qquad (9)$$

Simulation

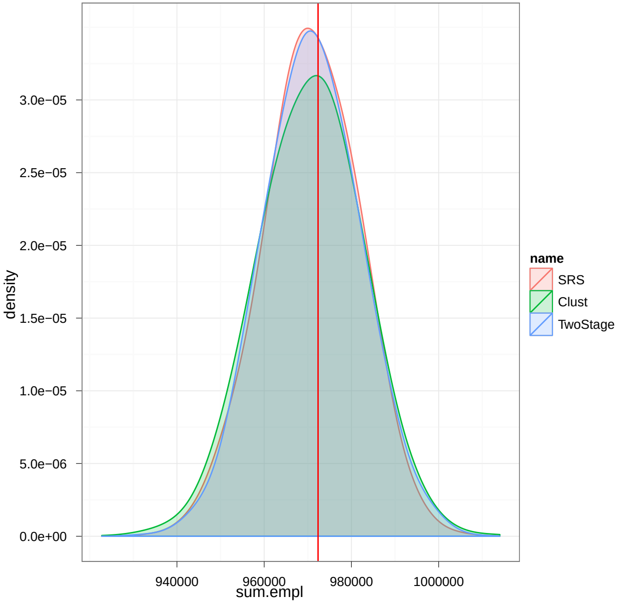
- Phase I: Find sample size for each sampling design so C_p is approximately equal for all sampling designs

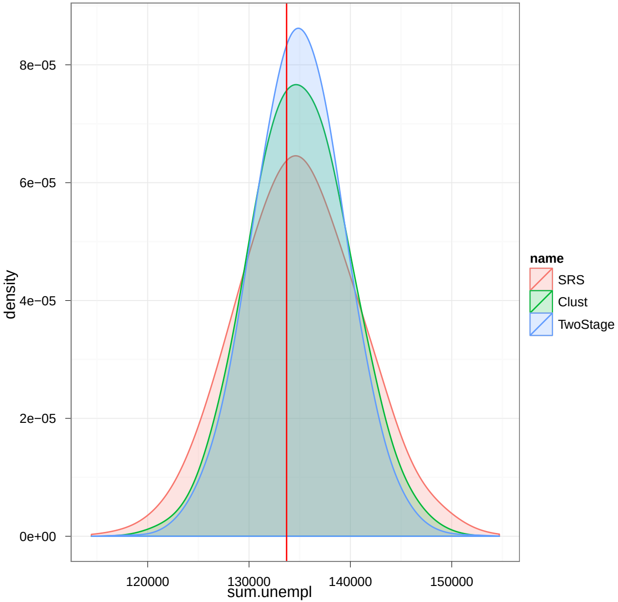


time.mean
14000
12000
10000
8000
6000
4000
2000
5000
10000
15000
20000
count.P.mean
design
SRS
Clust
TwoStage
time.sd
15
20
25
30
35

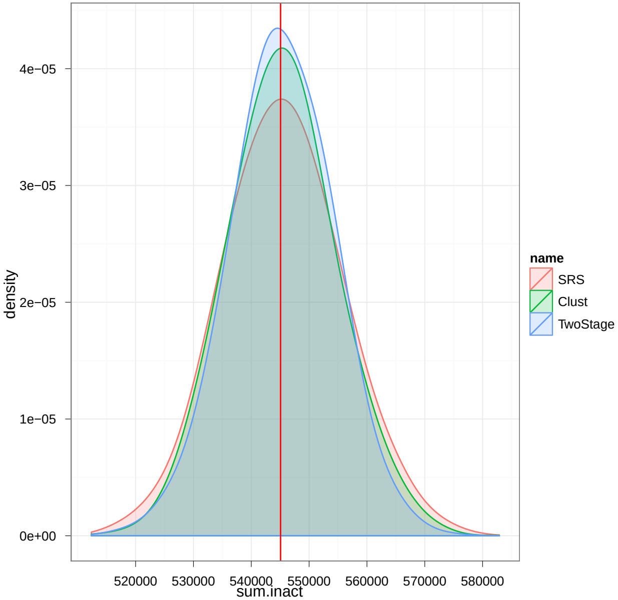
Simulation

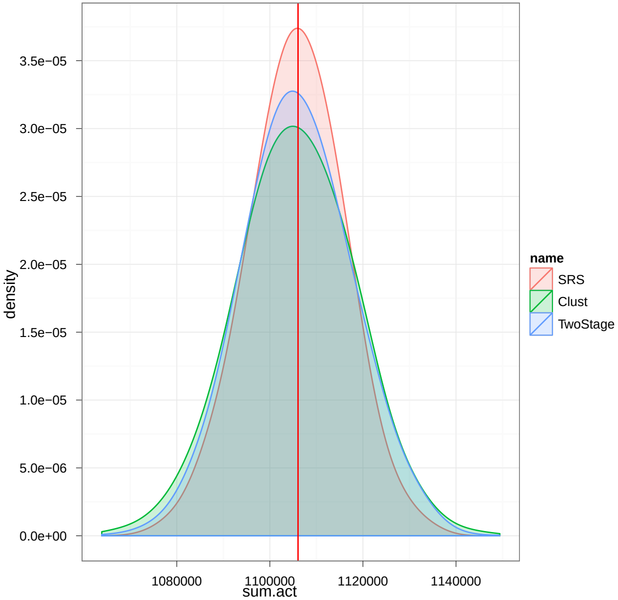
- Phase II: Estimate $V\left(\hat{\theta}_p\right)$ for each sampling design with sample size from Phase I


3.0e−05
2.5e−05
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
density
940000
960000
980000
1000000
sum.empl
name
SRS
Clust
TwoStage

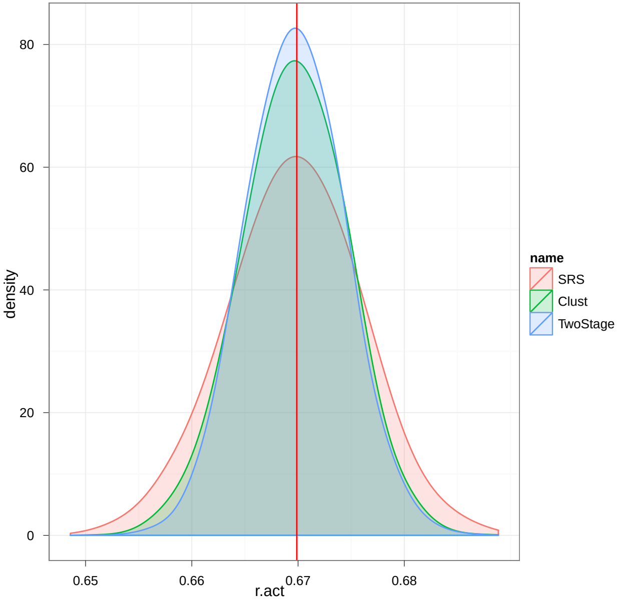


8e−05
6e−05
4e−05
2e−05
0e+00
density
120000
130000
140000
150000
sum.unempl
name
SRS
Clust
TwoStage

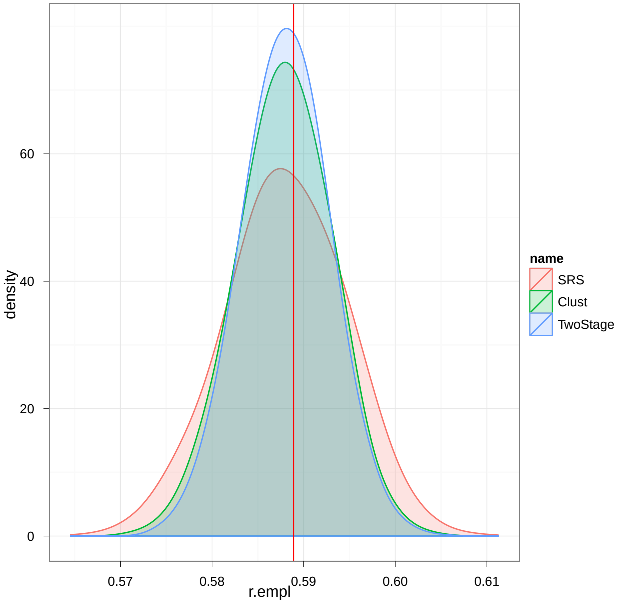

4e−05
3e−05
2e−05
1e−05
0e+00
density
520000
530000
540000
550000
560000
570000
580000
sum.inact
name
SRS
Clust
TwoStage



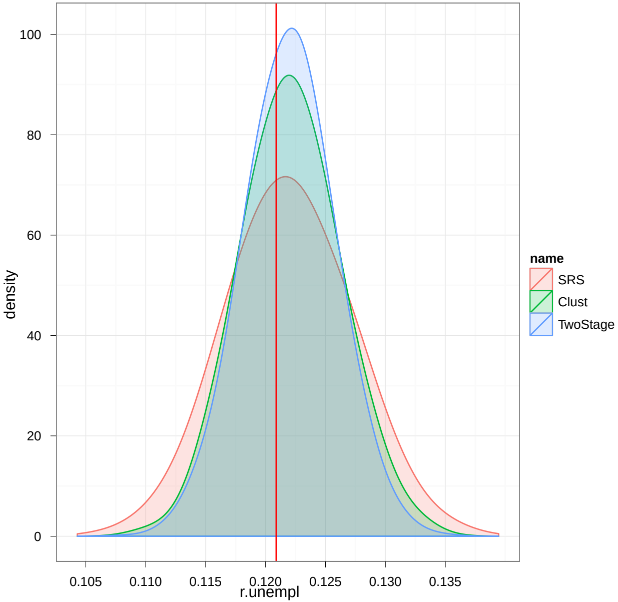
3.5e−05
3.0e−05
2.5e−05
2.0e−05
1.5e−05
1.0e−05
5.0e−06
0.0e+00
density
1080000
1100000
1120000
1140000
sum.act
name
SRS
Clust
TwoStage



80
60
40
20
0
density
0.65
0.66
0.67
0.68
r.act
name
SRS
Clust
TwoStage



density
60
40
20
0
0.57
0.58
0.59
0.60
0.61
r.empl
name
SRS
Clust
TwoStage


density
r.unempl
100
80
60
40
20
0
0.105
0.110
0.115
0.120
0.125
0.130
0.135
name
SRS
Clust
TwoStage

Outline

Conclusions

- The measure for design cost efficiency is introduced for one parameter of interest
- The measure for design cost efficiency for multi-parameter situation is necessary
- The tool (using simulations) for measuring design cost efficiency is under development

Thank you for your attention!