Eliciting Sensitive Data by Indirect Questioning Techniques: Some Recent Applications and Methodological Advances

Perri Pier Francesco

DESF, University of Calabria

CESS2016 - Conference of European Statistics Stakeholder
Budapest, 20-21 October 2016

Work partially supported by project PRIN-SURWEY (grant 2012F42NS8, Italy)

Outline

Doing sensitive research

In “**sensitive research**” on stigmatizing, highly personal, embarrassing, threatening or even incriminating issues, refusal to answer and misreporting represent **nonsampling errors** that are difficult to deal with and can lead to seriously biased analyses.
Although these errors cannot be totally avoided, they may be mitigated by increasing respondent cooperation.

Survey modes which ensure anonymity may improve confidentiality and, consequently, ensure more reliable information

Doing sensitive research

In "**sensitive research**" on stigmatizing, highly personal, embarrassing, threatening or even incriminating issues, refusal to answer and misreporting represent **nonsampling errors** that are difficult to deal with and can lead to seriously biased analyses.
Although these errors cannot be totally avoided, they may be mitigated by increasing respondent cooperation.

Survey modes which ensure anonymity may improve confidentiality and, consequently, ensure more reliable information

Beside traditional solutions (SAQs, CATI, CASI, CAWI, etc.), since the 1960s many different questioning methods have been devised to **ensure respondent anonymity and cutting down false reporting**

Indirect Questioning Techniques (IQTs)
(Chaudhuri and Christofides, 2013)

Indirect Questioning Techniques

IQTs include different approaches:

- the randomized response theory - RRT
- the non-randomized response technique - NRRT
- the item count technique - ICT
- the item sum technique - IST
- the nominative technique
- the three-card method
- ...

Indirect Questioning Techniques

IQTs include different approaches:

- the randomized response theory - RRT
- the non-randomized response technique - NRRT
- the item count technique - ICT
- the item sum technique - IST
- the nominative technique
- the three-card method
- ...

We focus on RRT, NRRT and IST

1. Results of some RRT/NRRT surveys
2. Improving the efficiency of IST estimates

Voluntary abortion in Italy

Official statistics show that, despite a slight reduction in last years of the number of voluntary abortions in Italy, the share of those made by foreigners is still growing

Gross abortion rate	
Italians	Foreigners
5.69	**26.73**

Source: Istat, 2011

If collecting data about abortion is difficult, obtaining data about illegal abortion is even more complicated. In Italy, the 69.3% of gynecologists refuse to put in practice abortion (for religious, ethical or other reasons), with a direct impact on the level of recourse to illegal abortion.
For foreign women illegally present in Italy, estimation of illegal abortion is more difficult:

- higher difficulties to integrate
- refuse the use of health facilities

Voluntary abortion in Italy

Official statistics show that, despite a slight reduction in last years of the number of voluntary abortions in Italy, the share of those made by foreigners is still growing

Gross abortion rate

Italians	Foreigners
5.69	**26.73**

Source: Istat, 2011

If collecting data about abortion is difficult, obtaining data about illegal abortion is even more complicated. In Italy, the 69.3% of gynecologists refuse to put in practice abortion (for religious, ethical or other reasons), with a direct impact on the level of recourse to illegal abortion.
For foreign women illegally present in Italy, estimation of illegal abortion is more difficult:

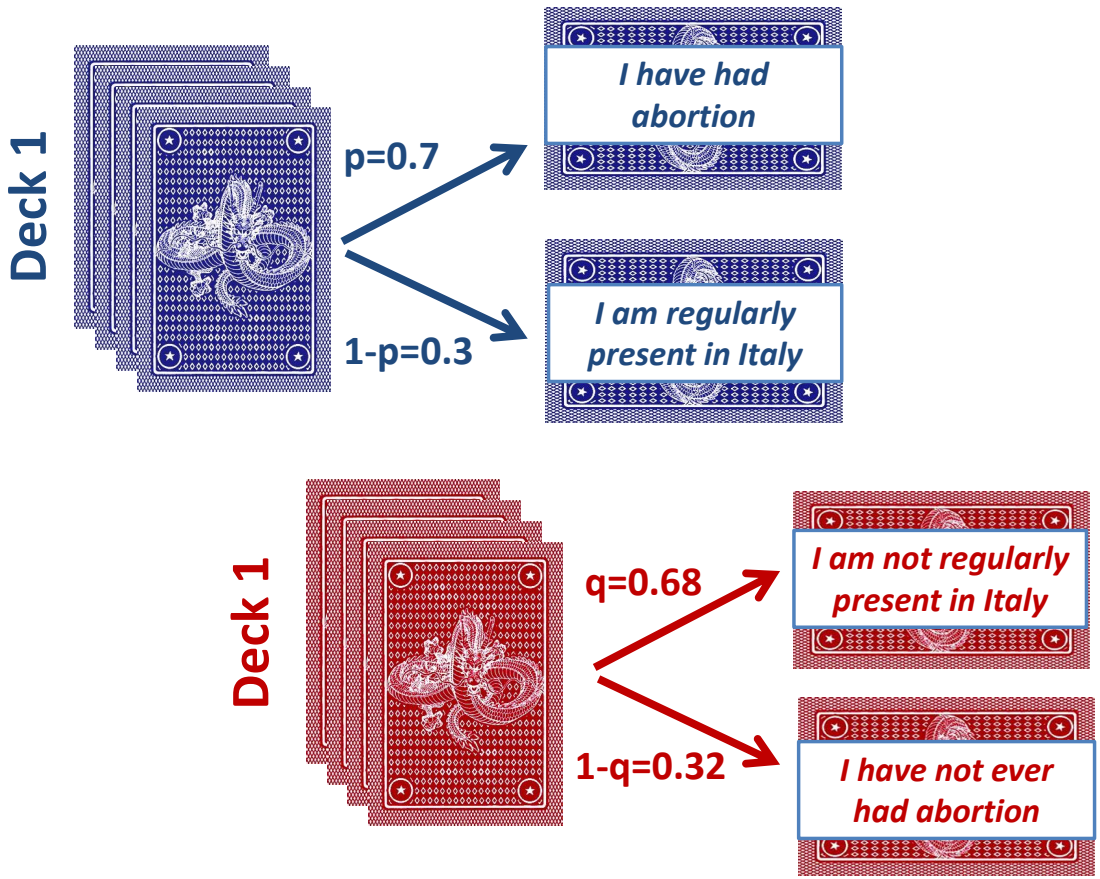
- higher difficulties to integrate
- refuse the use of health facilities

Motivated by these considerations, Perri et al. (2015) conducted a study to investigate a sensitive issue (induced abortion) in an elusive population (irregular immigrants)

The survey design

- A sample of 868 women spatially spread across the entire Calabria (south of Italy), with age average 31.8 years old and coming from 69 different countries, was considered (April 2014 - September 2014)
- The survey was administrated by face-to-face interviews conducted by thirty female graduate/final year students in Statistics at the University of Calabria
- Immigrant women were recruited by the interviewers via personal contacts in various aggregation points (religious places, leisure places, medical and assistance centers, phone centers, parks, squares, etc)
- Interviewees were firstly asked to provide socio-economic and demographic information through one-page short questionnaire. In 63.6% of cases, the interviewers compiled the questionnaire while, in the remainder, the women compiled it themselves
- Each woman was finally provided with a randomization device for collecting sensitive data on abortion and irregular presence

Crossed model (Lee et al., 2013)



- the randomized experiment was perfectly understood by 93.1% of the interviewees and correctly executed by 98%

Notation

- A: induced abortion (both legal and illegal)
- B: irregular presence of foreign women in Calabria
- π_A: prevalence of induced abortion
- π_B: prevalence of women illegally present in Calabria
- $\pi_{A \cap B}$: prevalence of women bearing both A and B
- $\pi_{A \cup B}$: prevalence of women bearing A or B or both

Notation

- A: induced abortion (both legal and illegal)
- B: irregular presence of foreign women in Calabria
- π_A: prevalence of induced abortion
- π_B: prevalence of women illegally present in Calabria
- $\pi_{A\cap B}$: prevalence of women bearing both A and B
- $\pi_{A\cup B}$: prevalence of women bearing A or B or both

Estimates of the unknown population parameters $\pi_A, \pi_B, \pi_{A\cap B}, \pi_{A\cup B}$ are obtained using the responses (Yes, Yes), (Yes, No), (No, Yes) and (No, No) collected in the sample.

As the distribution of the estimators is not normal, we used the *nonparametric bootstrap* and the *percentile method* to compute the 95% confidence interval.

Some results

	n	π_A	π_B	$\pi_{A\cap B}$	$\pi_{A\cup B}$
Sample	868	0.182	0.103	0.081	0.203
95%IC		[0.121 , 0.243]	[0.042 , 0.165]	[0.031 , 0.134]	[0.129 , 0.275]
Nationality					
Romanian	212	**0.393**	0.123	0.108	0.407
Other	656	0.114	0.096	0.073	0.137
Marital status					
Married/Cohabiting	361	0.198	0.124	0.070	0.252
Single	418	0.108	0.055	0.051	0.111
Separated/Divorced	76	0.485	0.202	0.293	0.396
Religion					
Catholic	281	0.163	0.135	0.112	0.185
Ortodox	254	0.427	0.107	0.121	0.412
Other	319	0.006	0.053	0.018	0.042
Employment status					
Working	471	0.222	0.113	0.096	0.239
Not Working	396	0.146	0.094	0.069	0.171
Contraception					
Yes	326	0.150	0.092	0.076	0.165
No	533	0.195	0.109	0.087	0.217

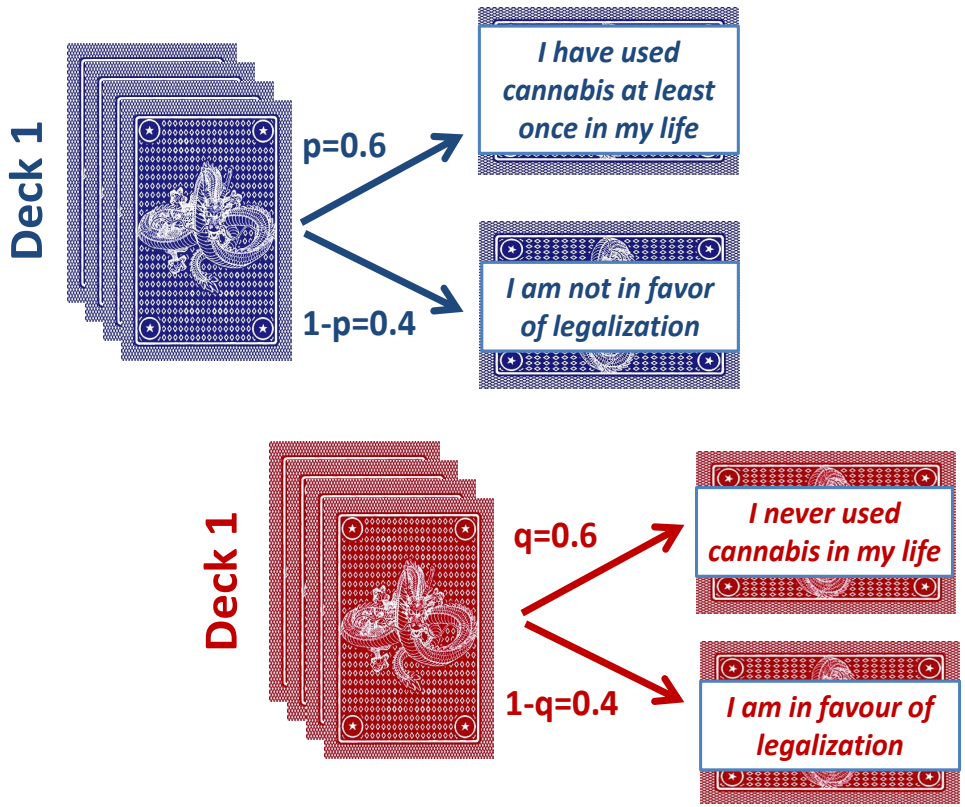
Some results

	n	π_A	π_B	$\pi_{A\cap B}$	$\pi_{A\cup B}$
Sample	868	0.182	0.103	0.081	0.203
95%IC		[0.121 , 0.243]	[0.042 , 0.165]	[0.031 , 0.134]	[0.129 , 0.275]
Nationality					
Romanian	212	**0.393**	0.123	0.108	0.407
Other	656	0.114	0.096	0.073	0.137
Marital status					
Married/Cohabiting	361	0.198	0.124	0.070	0.252
Single	418	0.108	0.055	0.051	0.111
Separated/Divorced	76	0.485	0.202	0.293	0.396
Religion					
Catholic	281	0.163	0.135	0.112	0.185
Ortodox	254	0.427	0.107	0.121	0.412
Other	319	0.006	0.053	0.018	0.042
Employment status					
Working	471	0.222	0.113	0.096	0.239
Not Working	396	0.146	0.094	0.069	0.171
Contraception					
Yes	326	0.150	0.092	0.076	0.165
No	533	0.195	0.109	0.087	0.217

- Official statistics (Istat, 2014) - based on hospital dismissal data - estimate the incidence of abortion among Romanians at **22.7%**

Cannabis use (A) and cannabis legalization (B)

- Cannabis use is more stigmatizing than cannabis legalization
- Perri et al. (2016) compared DQ method and the CM to investigate these two topics by a small survey conducted by face-to-face interviews in the city of Santa Maria del Cedro (CS)



Some results

	n	Method	π_A	π_B	$\pi_{A\cap B}$	$\pi_{A\cup B}$
Sample	289	CM	**0.471**	**0.686**	0.381	0.776
		DQ	**0.280**	**0.654**	0.253	0.681
Sex						
Women	135	CM	0.430	0.793	0.396	0.826
		DQ	0.170	0.585	0.163	0.592
Men	154	CM	0.507	0.591	0.367	0.730
		DQ	0.3770	0.714	0.331	0.76

According to the "more-is-better assumption", the CM seems to work better than the DQ

Some benchmark

- In 2014, **31.9%** was the prevalence of cannabis users in Italy aged 15$\geq$ (*Source: European Monitoring Center for Drugs and Drug Addition*)
- **73%** of people aged 18$\geq$ support cannabis legalization (*Source: Ipsos Public Affairs*)

Racism at University of Calabria

- Italian and foreign students of University of Calabria have been invited to fill in an online questionnaire
- Personal contacts and social networks (facebook, whatsapp,...) have be used to invite students

Racism at University of Calabria

- Italian and foreign students of University of Calabria have been invited to fill in an online questionnaire
- Personal contacts and social networks (facebook, whatsapp,...) have be used to invite students

Racism is...

Believing that some races are superior to others is the cornerstone of racist thinking. If you believe that the race to which you belong (or one where you do not belong) possess the qualities that make it better than others, this is a racist thinking.

Two sensitive direct questions

- **Question 1**: Do you usually have any negative attitude/behavior towards the races/ethnic group? (*in the middle of the questionnaire*)
- **Question 2**: Do you think you are a **racist** against people different to your race/ethnic group? (*at the end of the questionnaire*)

Crosswise/triangular models (Tian & Tang, 2014)

Instructions for the respondents: "*Please read carefully the four statements indicated below and find the one that reflects your status with respect to your birthday and racism*"

A *My birthday is between October and December and I am a racist*

B *My birthday is between January and September and I am a racist*

C *My birthday is between October and December and I am not a racist*

D *My birthday is between January and September and I am not a racist*

Crosswise/triangular models (Tian & Tang, 2014)

Instructions for the respondents: "*Please read carefully the four statements indicated below and find the one that reflects your status with respect to your birthday and racism*"

A *My birthday is between October and December and I am a racist*

B *My birthday is between January and September and I am a racist*

C *My birthday is between October and December and I am not a racist*

D *My birthday is between January and September and I am not a racist*

Crosswise model (CWM)

- If your status fits statement A or D, please put a tick here ◯
- If your status fits statement B or C, please put a tick here ◯

Crosswise/triangular models (Tian & Tang, 2014)

Instructions for the respondents: "*Please read carefully the four statements indicated below and find the one that reflects your status with respect to your birthday and racism*"

A *My birthday is between October and December and I am a racist*

B *My birthday is between January and September and I am a racist*

C *My birthday is between October and December and I am not a racist*

D *My birthday is between January and September and I am not a racist*

Crosswise model (CWM)

- If your status fits statement A or D, please put a tick here ◯
- If your status fits statement B or C, please put a tick here ◯

Triangular model (TM)

- If your status fits statement A, B or C, please put a tick here ◯
- If your status fits statement D, please put a tick here ◯

Crosswise/triangular models (Tian & Tang, 2014)

Instructions for the respondents: "*Please read carefully the four statements indicated below and find the one that reflects your status with respect to your birthday and racism*"

A *My birthday is between October and December and I am a racist*

B *My birthday is between January and September and I am a racist*

C *My birthday is between October and December and I am not a racist*

D *My birthday is between January and September and I am not a racist*

Crosswise model (CWM)

- If your status fits statement A or D, please put a tick here ◯
- If your status fits statement B or C, please put a tick here ◯

Triangular model (TM)

- If your status fits statement A, B or C, please put a tick here ◯
- If your status fits statement D, please put a tick here ◯

In both the models, privacy is protected

Some preliminary results

		DQ		NRRT	
	n	Q1	Q2	CWM	TM
Sample	567	0.18	0.13	**0.22**	**0.27**
Sex					
Women	297	0.18	0.14	**0.14**	**0.23**
Men	270	0.18	0.11	**0.30**	**0.31**

Some preliminary results

		DQ		NRRT	
	n	Q1	Q2	CWM	TM
Sample	567	0.18	0.13	**0.22**	**0.27**
Sex					
Women	297	0.18	0.14	**0.14**	**0.23**
Men	270	0.18	0.11	**0.30**	**0.31**

- the proportion of affirmative responses to Q1 is slightly higher than to Q2
- CWM and TM provide estimates which are higher than those under DQ

According to the "more-is-better assumption", the TM seems to work better than the CWM

The Item Sum Technique

The IST is a variant of the ICT, suitable for **quantitative** sensitive characteristics, introduced by Chaudhuri and Christofides (2013) and firstly used by Trappmann et al. (2014) in a CATI survey on undeclared work in Germany.

The Item Sum Technique

The IST is a variant of the ICT, suitable for **quantitative** sensitive characteristics, introduced by Chaudhuri and Christofides (2013) and firstly used by Trappmann et al. (2014) in a CATI survey on undeclared work in Germany.

Procedure

1. two independent samples, s_1 and s_2, are drawn from the population
2. units in s_1 are presented with a long list (LL) of items containing $(G+1)$ questions, G of these are innocuous and one is sensitive. Units in s_2 receive a short list (SL) containing only the G innocuous questions
3. all the items refer to quantitative variables possibly measured on the same scale of the sensitive one
4. the respondents are asked to report the **total score** of the answers to all the questions in their list without revealing the individual score of each question
5. the mean difference of answers between the LL-sample and the SL-sample is then used as an unbiased estimator of the population mean of the sensitive variable

Notation

In the IST estimation, Rueda et al. (2016) proposed some methodological advances. Let:

- $U = \{1, \ldots, N\}$ is a finite population of N units
- y_i is the value of the sensitive variable, y, for unit $i \in U$
- $\bar{Y} = N^{-1} \sum_{i \in U}^{N} y_i$ is the unknown mean to be estimated
- $p(\cdot)$ is generic sampling design with first-order inclusion probabilities π_i
- $d_i = \pi_i^{-1}$ is the **sampling design-basic weight** for the i-unit
- s_1 (LL-sample) and s_2 (SL-sample) are two independent samples selected from U according to $p(\cdot)$
- t is the variable denoting the total score applicable to the G nonsensitive questions $\rightarrow$ SL-sample
- $z = y + t$ is the total score applicable to the nonsensitive questions and the sensitive question $\rightarrow$ LL-sample

... the answer of the ith respondent will be $z_i = y_i + t_i$ if $i \in s_1$ or t_i if $i \in s_2$

Notation

In the IST estimation, Rueda et al. (2016) proposed some methodological advances. Let:

- $U = \{1, \ldots, N\}$ is a finite population of N units
- y_i is the value of the sensitive variable, y, for unit $i \in U$
- $\bar{Y} = N^{-1} \sum_{i \in U}^{N} y_i$ is the unknown mean to be estimated
- $p(\cdot)$ is generic sampling design with first-order inclusion probabilities π_i
- $d_i = \pi_i^{-1}$ is the **sampling design-basic weight** for the i-unit
- s_1 (LL-sample) and s_2 (SL-sample) are two independent samples selected from U according to $p(\cdot)$
- t is the variable denoting the total score applicable to the G nonsensitive questions $\rightarrow$ SL-sample
- $z = y + t$ is the total score applicable to the nonsensitive questions and the sensitive question $\rightarrow$ LL-sample

... the answer of the ith respondent will be $\mathbf{z_i = y_i + t_i}$ if $\mathbf{i \in s_1}$
or $\mathbf{t_i}$ if $\mathbf{i \in s_2}$

Horvitz-Thompson estimation

Under the design $p(\cdot)$, let

$$\hat{\bar{Z}}_{HT} = \frac{1}{N}\sum_{i\in s_1} d_i z_i \quad \text{and} \quad \hat{\bar{T}}_{HT} = \frac{1}{N}\sum_{i\in s_2} d_i t_i$$

be the unbiased Horvitz-Thompson (hereafter HT) estimators of $\bar{Z} = N^{-1}\sum_{i\in U}(y_i + t_i)$ and $\bar{T} = N^{-1}\sum_{i\in U} t_i$

Hence, an unbiased **HT-type estimator** of $\bar{Y}$ can be immediately obtained as

$$\hat{\bar{Y}}_{HT} = \hat{\bar{Z}}_{HT} - \hat{\bar{T}}_{HT}$$

Calibration estimation

In sampling practice, if auxiliary information is available, estimates can be improved by **calibration** (Deville and Särndal, 1992; Särndal, 2007). Let:

- $\mathbf{x}$ a vector of k auxiliary variables
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ the vector of the **known** values for unit $i \in U$
- $\bar{\mathbf{X}} = N^{-1} \sum_{i \in U} \mathbf{x}_i$ the vector for the known population means of the k auxiliary variables
- ω_{ij} new weights based on sample s_j $(j = 1, 2)$ obtained by minimizing

$$\Phi_{s_j}(d_i, \omega_{ij}) = \sum_{i \in s_j} \frac{(\omega_{ij} - d_i)^2}{d_i q_i}, \quad j = 1, 2$$

subject to the **calibration equations** $N^{-1} \sum_{i \in s_j} \omega_{ij} \mathbf{x}_i = \bar{\mathbf{X}}$.

Calibration estimation

In sampling practice, if auxiliary information is available, estimates can be improved by **calibration** (Deville and Särndal, 1992; Särndal, 2007). Let:

- $\mathbf{x}$ a vector of k auxiliary variables
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ the vector of the **known** values for unit $i \in U$
- $\bar{\mathbf{X}} = N^{-1} \sum_{i \in U} \mathbf{x}_i$ the vector for the known population means of the k auxiliary variables
- ω_{ij} new weights based on sample s_j $(j = 1, 2)$ obtained by minimizing

$$\Phi_{s_j}(d_i, \omega_{ij}) = \sum_{i \in s_j} \frac{(\omega_{ij} - d_i)^2}{d_i q_i}, \quad j = 1, 2$$

subject to the **calibration equations** $N^{-1} \sum_{i \in s_j} \omega_{ij} \mathbf{x}_i = \bar{\mathbf{X}}$.

According to the $w'_{ij}s$, we define the **calibration estimator** of $\bar{Y}$ as

$$\hat{\bar{Y}}_C = \hat{\bar{Z}}_C - \hat{\bar{T}}_C$$

where

$$\hat{\bar{Z}}_C = \frac{1}{N} \sum_{i \in s_1} \omega_{i1} z_i \quad \text{and} \quad \hat{\bar{T}}_C = \frac{1}{N} \sum_{i \in s_2} \omega_{i2} t_i$$

Simulation data

A simulation study has been conducted to investigated the performance of the HT and calibration estimators $(\hat{\bar{Y}}_{HT}, \hat{\bar{Y}}_{C})$ when data are supposed to be collected by the IST.
The study is based on the "World Bank's Enterprise Surveys" data collected in China in 2011:

- we considered the surveyed sample of 2848 firms as the target population
- **study variable:** the fiscal total sales for year 2011
- **innocuous variable for IST:** the total annual costs of electricity
- **calibrating variables:** total sales for year 2009, full-time workers for year 2009, and firm's yearly average inventories in finished goods in 2011
- to evaluate the performance of estimation under the IST, we also considered the HT and calibration estimators $(\hat{\bar{Y}}_{HTy}, \hat{\bar{Y}}_{Cy})$ computed without IST only on the basis of the true value of Y

Adopted sampling designs

Data at firm-level have been conceived as the target population from which samples of size n between 25 and 200 firms have been selected according to:

1. **simple random sampling without replacement** (SRSWOR)
2. **stratified SRSWOR**
   Population has been stratified into three industrial sectors:"manufacturing", "retail" and "other services". From each stratum, samples have been selected according to SRSWOR with proportional allocation from 5% to 15% of the population size
3. **Midzuno sampling design**
   The sampling design has been implemented with first-order inclusion probabilities proportional to the number of establishments that form the firm

Accuracy measures

The performance of the HT estimator and calibration estimator, under IST ($\rightarrow \hat{\bar{Y}}^* = \hat{\bar{Y}}_{HT}, \hat{\bar{Y}}_C$) and DQ ($\rightarrow \hat{\bar{Y}}_{HTy}, \hat{\bar{Y}}_{Cy}$), has been evaluated by means of the absolute relative bias

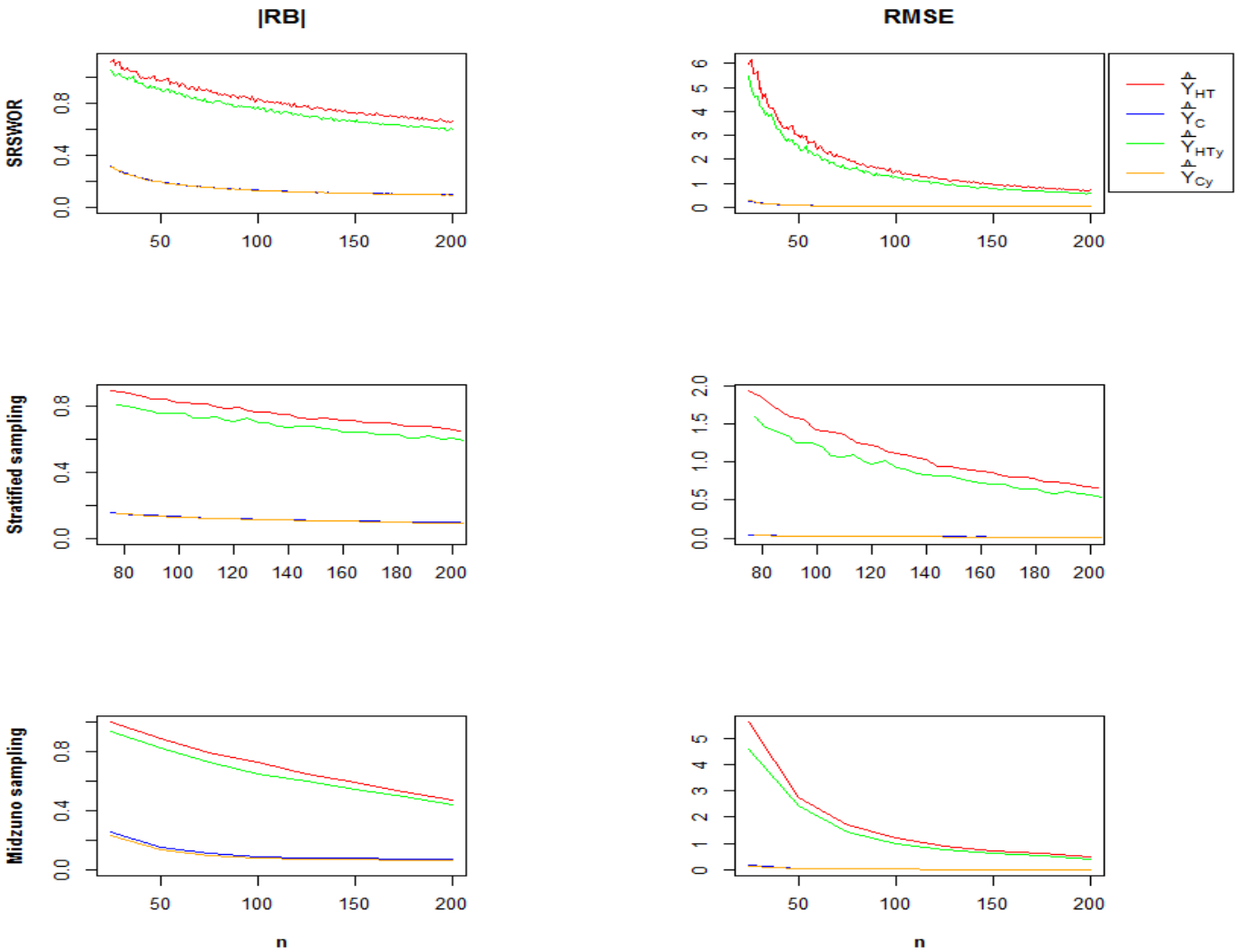
$$|\text{RB}(\hat{\bar{Y}}^*)| = \left| \frac{\mathbb{E}_M(\hat{\bar{Y}}^*) - \bar{Y}}{\bar{Y}} \right|$$

and the relative mean square error for the estimator

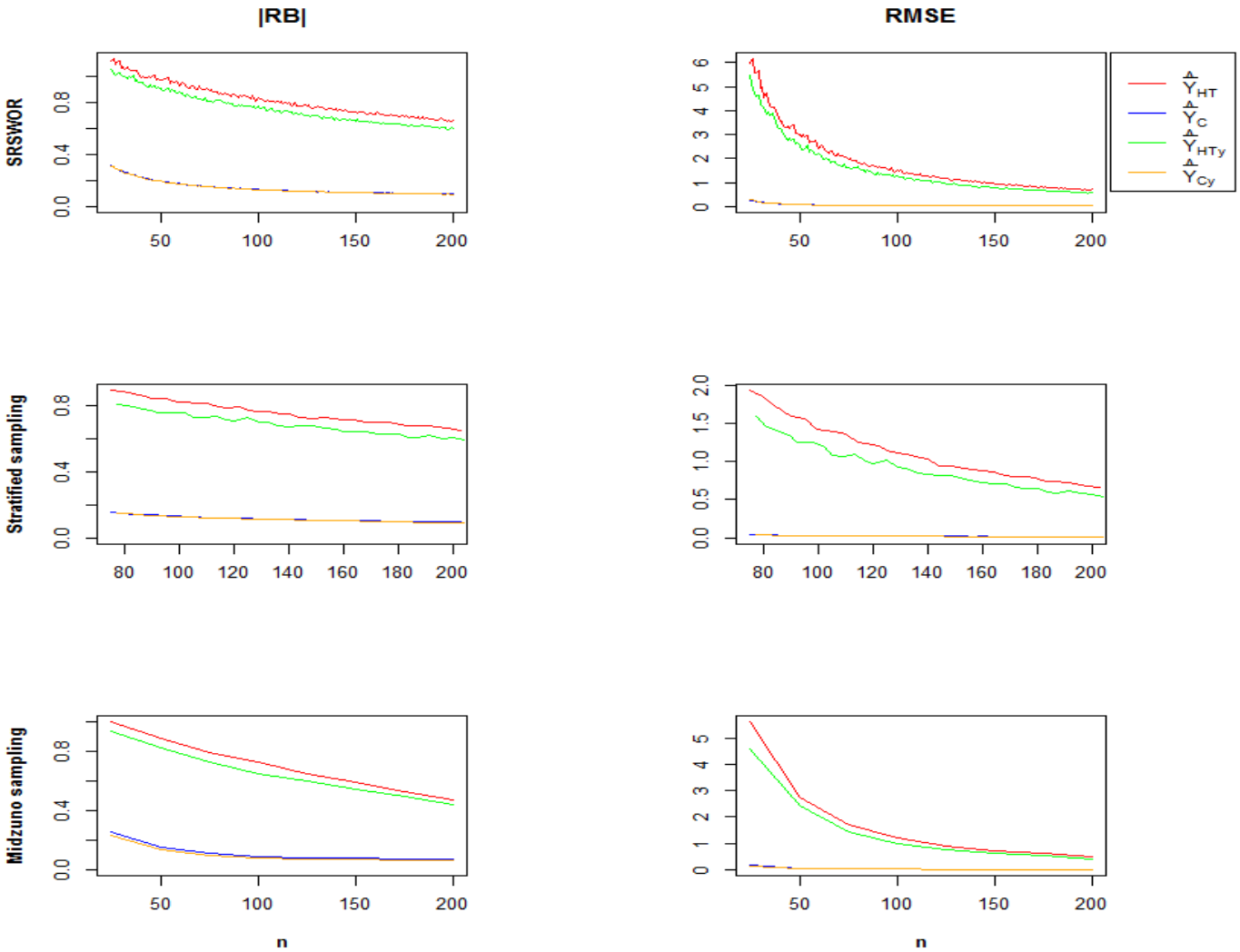
$$\text{RMSE}(\hat{\bar{Y}}^*) = \frac{\mathbb{E}_M(\hat{\bar{Y}}^* - \bar{Y})^2}{\bar{Y}^2}$$

where the operator $\mathbb{E}_M$ is evaluated on the basis of 10,000 Monte Carlo replications.

Results

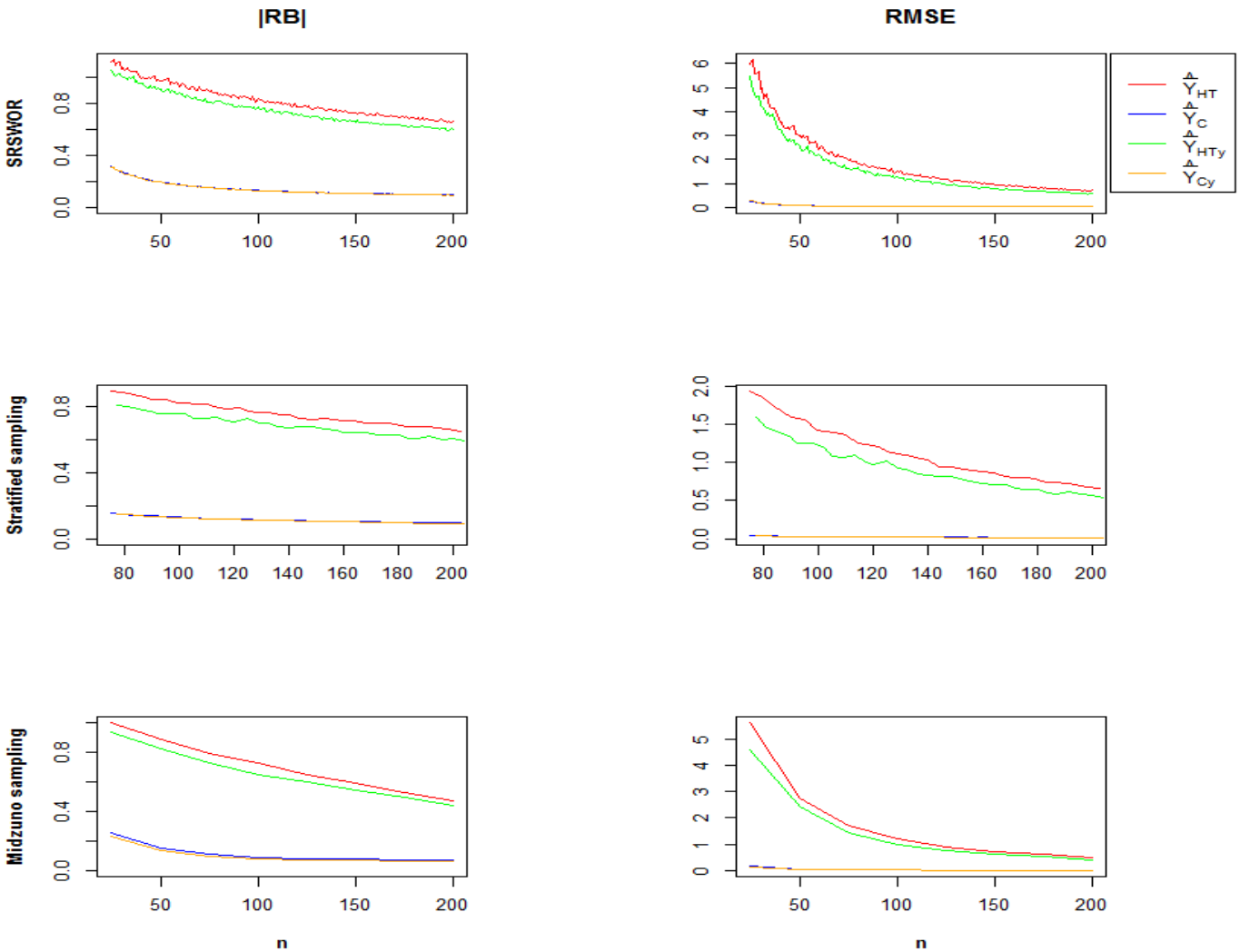

|RB|
RMSE
SRSWOR
Stratified sampling
Midzuno sampling
n
n
$\hat{Y}_{HT}$
$\hat{Y}_{C}$
$\hat{Y}_{HTy}$
$\hat{Y}_{Cy}$

Results



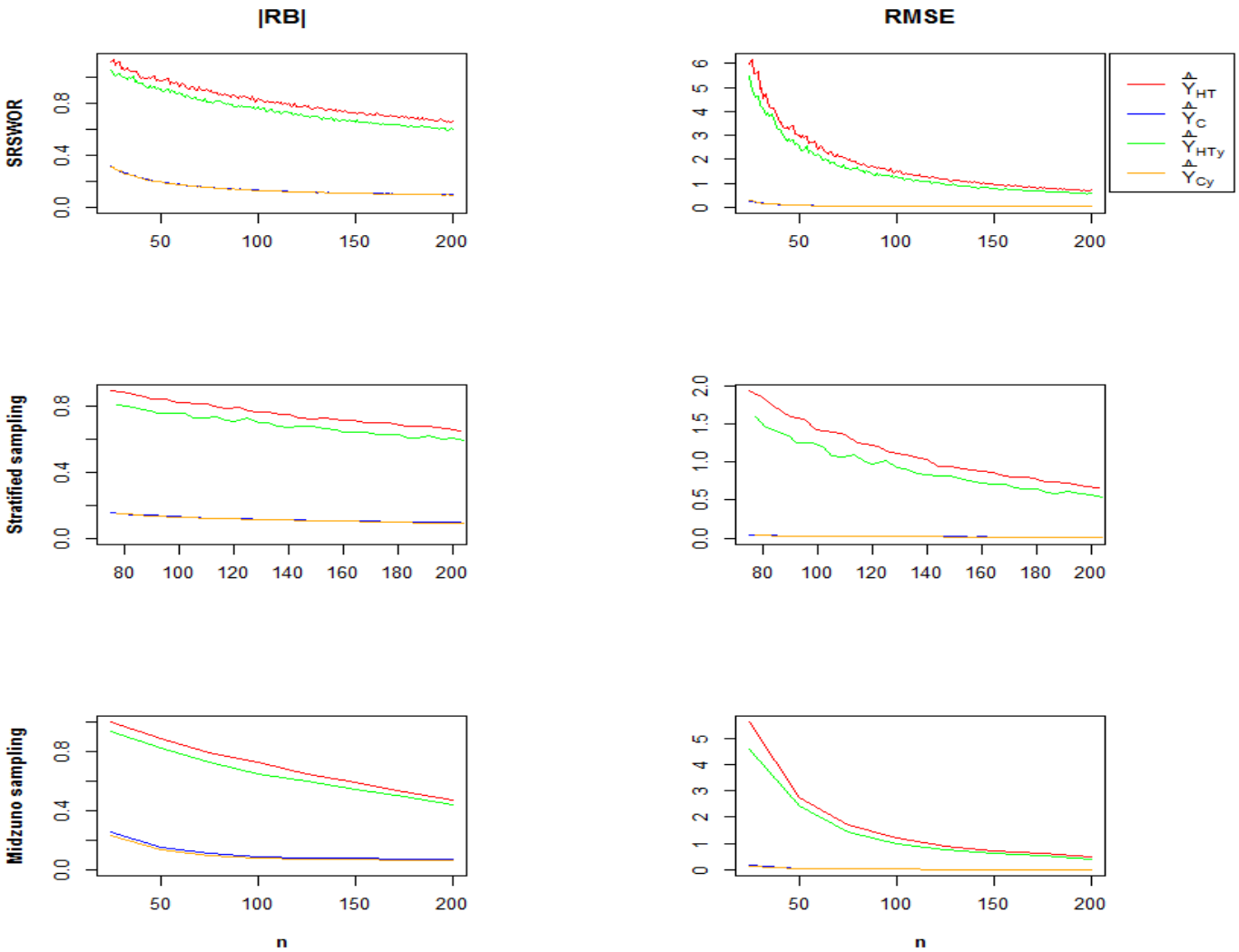
The HT estimator based on the y_i's slightly outperforms, as expected, the HT estimator based on the IST values z_i

Results



The calibration estimators are unexpectedly nearly equivalent both in term of (absolute) bias and mean square error

Results



Using auxiliary information through stratification and sampling with varying probabilities improves the efficiency of the estimates w.r.t. SRSWOR

Thank you for your attention

Department of Economics, Statistics and Finance
University of Calabria (Italy)

pierfrancesco.perri@unical.it

References

1. Chaudhuri A., Christofides T.C. (2013). *Indirect Questioning in Sample Surveys.* Springer-Verlag Berlin Heidelberg
2. Deville J.C., and Särndal C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382
3. Istat (2014a). Induced abortion, Year 2012. Available at http://www.istat.it/en/archive/141810
4. Lee C.S., Sedory S.A., Singh S. (2013). Estimating at least seven measures of qualitative variables from a single sample using randomized response technique. *Statistics and Probability Letters*, 83, 399-409
5. Perri P.F., Pelle E., Stranges M. (2015) Estimating induced abortion and foreign irregular presence using the randomized response crossed model. *Social Indicators Research*, online first, DOI: 10.1007/s11205-015-1136-x
6. Perri P.F., Pelle E., Aloise G. (2016) Eliciting sensitive data via the randomized response model: some evidence from a study on the illicit cannabis use and its legalization. *Work in progress*
7. Perri P.F., Trang V. Are students of University of Calabria racist? Some evidences from triangular and crosswise nonrandomized response models in a Web-survey. *Work in progress*
8. Rueda M., Perri P.F., Cobo B.R. (2016) Advances in estimation by the item sum technique using auxiliary information in complex surveys. Submitted article
9. Särndal C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119
10. Tian G.-L., Tang M.-L. (2014). *Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys.* Chapman & Hall/CRC, Boca Raton, FL
11. Trappmann M., Krumpal I., Kirchner A., Jann B. (2014). Item sum: A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology*, 2, 58-77