Preserving privacy protection using indirect questioning techniques in real sensitive surveys

Perri Pier Francesco

DESF, University of Calabria

CESS2018 - Conference of European Statistics Stakeholder
Bamberg, 18-19 October 2018

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
- ...

In this talk focus on

- **RRT and IST**
- **Design and results of some sensitive surveys**

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 and ethical), 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 and ethical), 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

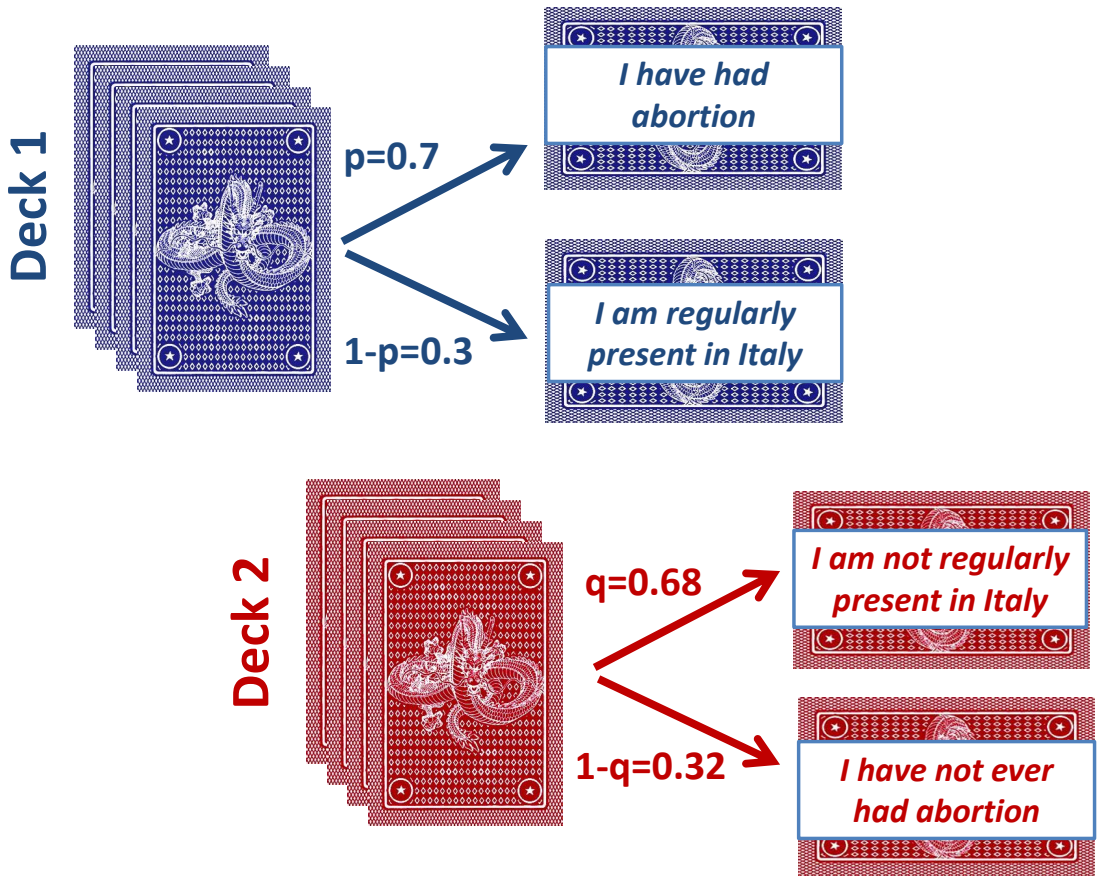
- higher difficulties to integrate
- refuse the use of health facilities

Motivated by these considerations, Perri et al. (2016) 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 to collect 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

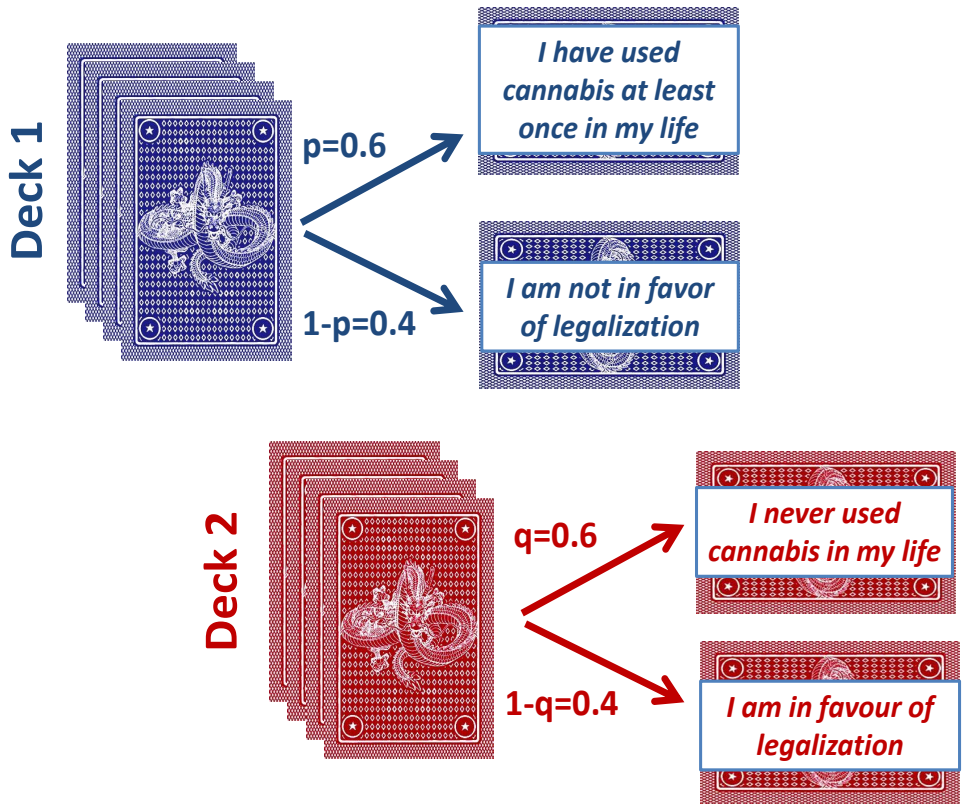
Some results

	n	π_A	π_B	$\pi_{A\cap B}$	$\pi_{A\cup B}$
Sample	868	0.182	0.103	0.081	0.203
95%CI		[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 and Pelle (2018) compared DQ method and the CM to investigate these two topics in a small survey conducted by face-to-face interviews ($n = 289$) in municipality located in southern Italy



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*)

The Item Sum Technique

The IST is a variant of the ICT, suitable for **quantitative** sensitive characteristics, firstly introduced and used by Trappmann et al. (2011, 2014)

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

Some (methodological) advances in the IST

1. Rueda et al. (2018) discussed the profitable use of auxiliary information at the estimation stage through the calibration approach
2. Perri et al. (2018a) discussed the
   - implementation of the IST when two or more sensitive variables are investigated and their means are to be estimated
   - determination of the optimal sample size for the SL and LL groups to achieve minimum variance estimates
3. Perri et al. (2018b) used the IST in a real survey among students in Spain

Cannabis use & sexual addiction

- A mixed-mode survey was conducted at the universities of Granada and Murcia (Spain) during the academic year 2015/2016
- A stratified sample of 2398 students enrolled in different faculties was considered
- Three survey modes to collect sensitive information were used: DQ, IST and Bar-Lev et al. (2004) RR method
- Students were contacted in class and randomly assigned to one of the three survey modes. A questionnaire was distributed during the class time break

Cannabis use & sexual addiction

- A mixed-mode survey was conducted at the universities of Granada and Murcia (Spain) during the academic year 2015/2016
- A stratified sample of 2398 students enrolled in different faculties was considered
- Three survey modes to collect sensitive information were used: DQ, IST and Bar-Lev et al. (2004) RR method
- Students were contacted in class and randomly assigned to one of the three survey modes. A questionnaire was distributed during the class time break

DQ survey mode, $n = 492$ students (20.5%)

Q1: How many cannabis cigarettes did you consume last year?

Q2: Over the past 90 days, how many days did you consume cannabis?

Q3: Over the past 90 days, how many times have you had trouble stopping your sexual behaviour when you knew it was inappropriate?

Q4: Over the past 90 days, how many times has sex been an escape from your problems?

Cannabis use & sexual addiction

IST survey, 773 (32.2%) students in the LL-sample and 520 (17.7%) in the SL-sample

- IST was repeated four times, one run for each of the sensitive questions Q1-Q4
- Four different nonsensitive, innocuous, questions were formulated
- Students in the LL-sample received a questionnaire with text explaining the IST procedure followed by a block consisting of pairs of questions: the sensitive question and the corresponding nonsensitive question

Cannabis use & sexual addiction

IST survey, 773 (32.2%) students in the LL-sample and 520 (17.7%) in the SL-sample

- IST was repeated four times, one run for each of the sensitive questions Q1-Q4
- Four different nonsensitive, innocuous, questions were formulated
- Students in the LL-sample received a questionnaire with text explaining the IST procedure followed by a block consisting of pairs of questions: the sensitive question and the corresponding nonsensitive question

LL-sample	SL-sample
Q1: How many cannabis cigarettes did you consume last year? **IQ1:** What was your general mark in the Selectivity exam, without counting specific subjects? (Value between 0 and 10)	**IQ1**
report the sum of the scores of the two questions, without revealing the individual responses	
Q2: Over the past 90 days, how many days did you consume cannabis? **IQ2:** What was your Selectivity mark counting specific subjects? (Value between 0 and 14)	**IQ2**
report the sum of the scores of the two questions, without revealing the individual responses	
Q3: Over the past 90 days, how many times have you had trouble stopping your sexual behaviour when you knew it was inappropriate? **IQ3:** What is the number of subjects in which you have enrolled in the academic year?	**IQ3**
report the sum of the scores of the two questions, without revealing the individual responses	
Q4: Over the past 90 days, how many times has sex been an escape from your problems? **IQ4:** What is the final digit of your mobile phone number?	**IQ4**
report the sum of the scores of the two questions, without revealing the individual responses	

Cannabis use & sexual addiction by BarLev method

With probability q the ith respondent is asked to release the true value $\mathbf{y_i}$ of the sensitive variable whereas with probability $1-q$ is asked to generate a number $\mathbf{s_i}$ from an innocuous variable and multiply it by y_i. Hence, the observed randomized response for the ith respondent is

$$z_i = \begin{cases} y_i & \text{with probability } q \\ y_i s_i & \text{with probability } 1-q \end{cases}$$

Cannabis use & sexual addiction by BarLev method

With probability q the ith respondent is asked to release the true value $\mathbf{y_i}$ of the sensitive variable whereas with probability $1-q$ is asked to generate a number $\mathbf{s_i}$ from an innocuous variable and multiply it by y_i. Hence, the observed randomized response for the ith respondent is

$$z_i = \begin{cases} y_i & \text{with probability } q \\ y_i s_i & \text{with probability } 1-q \end{cases}$$

BarLev RR survey, 613 students (26.6%)

As randomizing device we used the smartphone application of the "Baraja Española", a deck composed of 40 cards, divided into four families or suits, each numbered from 1 to 7, and three figures for the each suit

- Students were requested to install the application on their smartphone
- For each sensitive question Q1-Q4, students were asked to run the application and to give the true sensitive response y_i if the card shown was a figure ($q = 12/40$). If not a figure, students were asked to multiply y_i by the number shown on the card

Results: nonresponse rates

Nonresponse rates (%) for DQ, BarLev and IST survey modes

Question	DQ	BarLev	IST
Q1	10.96	14.03	$1.93^{**}_{\star\star\star}$
Q2	11.79	4.40^{*}	1.55^{**}
Q3	21.14	6.69^{**}	$0.15^{***}_{\star\star}$
Q4	16.67	6.20^{**}	$0.23^{***}_{\star\star}$

One-tailed *t*-test for difference between two proportions

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ for IQTs vs DQ

$^{\star}p < 0.05$, $^{\star\star}p < 0.01$, $^{\star\star\star}p < 0.001$ for IST vs BarLev

- DQ nonresponse rate is higher for questions Q3 and Q4 than for Q1 and Q2. This is probably due to the fact that sexual matters are much more confidential than are patterns of cannabis consumption
- A significant reduction in the nonresponse rate is observed in the case of the IQTs, particularly for Q3 which seems to be the most sensitive one
- The comparison between the two IQTs reveals that the IST nonresponse rate is statistically lower than that of the BarLev method. In general, the IST yielded a very low nonresponse rate, no more than 2% for any of questions Q1-Q4

Results: mean estimates 1/2

		DQ			BarLev			IST		
	Question	Mean	SD	95% CI	Mean	SD	95% CI	Mean	SD	95% CI
Total										
	Q1	3.11	0.60	[1.93;4.30]	13.07**	3.38	[6.43;19.70]	14.93***	2.53	[9.97;19.89]
	Q2	1.41	0.40	[0.63;2.19]	9.33***	1.28	[6.82;11.84]	3.72^{***}_{***}	0.47	[2.80;4.65]
	Q3	0.23	0.07	[0.10;0.36]	2.12***	0.42	[1.31;2.94]	1.11^{***}_{**}	0.29	[0.53;1.68]
	Q4	2.52	0.66	[1.23;3.81]	3.46	0.55	[2.38;4.53]	7.60^{***}_{***}	0.70	[6.24;8.97]
Males										
	Q1	6.35	1.43	[3.54;9.15]	21.14*	7.2	[7.03;35.25]	24.65***	4.69	[15.47;33.84]
	Q2	2.23	0.76	[0.74;3.72]	8.85***	1.67	[5.58;12.12]	5.51^{***}_{*}	0.81	[3.92;7.09]
	Q3	0.48	0.17	[0.15;0.81]	2.73**	0.90	[0.97;4.48]	1.94***	0.57	[0.82;3.07]
	Q4	3.98	1.26	[1.51;6.44]	3.65	0.91	[1.87;5.43]	8.16^{***}_{***}	1.05	[6.10;10.22]
Females										
	Q1	0.25	0.12	[0.01;0.49]	7.91**	3.06	[1.90;13.91]	6.48***	2.34	[1.89;11.06]
	Q2	0.82	0.49	[-0.14;1.78]	9.76***	1.85	[6.13;13.39]	2.17^{**}_{***}	0.52	[1.15;3.18]
	Q3	0.07	0.03	[0.02;0.12]	1.75***	0.37	[1.03;2.47]	0.39^{***}_{***}	0.17	[0.01;0.77]
	Q4	1.86	0.83	[0.23;3.50]	3.25	0.68	[1.91;4.60]	7.08^{***}_{***}	0.93	[5.26;8.91]

One-tailed t-test for differences in means.

$^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$ for IQTs vs DQ, and $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$ for IST vs BarLev

- DQ estimates are lower than the IQT ones
- BarLev estimates are higher than the IST ones for Q2 and Q3, and lower for Q4

According to the "more-is-better" assumption, both of the IQTs outperform the DQ method, but there is no evidence about a uniform superiority of one IQT over the other

- IST estimates present lower SD and narrower CI than the BarLev method, except for Q4. As expected, DQ estimates are more precise than the IQT ones

Results: mean estimates 2/2

Results show that

Patterns of cannabis consumption and sexual addiction are present among students, with a slight predominance in the male group

- BarLev method indicates that, on average, **2.12** times [0.23 DQ; 1.11 IST] students had difficulty in halting inappropriate sexual behaviour (2.73 M and 1.75 F)
- IST estimates suggest a more frequent use of sex to escape from personal problems, on average **7.6** times [2.52 DQ; 3.46 BarLev] (8.16 M and 7.08 F)
- According to IQTs, students smoked around **14** cannabis cigarettes [3.11 DQ]. In particular, under BarLev
  - male students smoked more cigarettes than female students (21.14 vs 7.91)
  - students on average consumed cannabis on **9.33** days [1.14 DQ; 3.72 IST] (8.85 M and 9.76 F)

A study on non-heterosexual identity

- Hsieh and Perri (2018) planned a pilot study on sexual identity in Taiwan by using the RR procedure proposed by Christofides' (2003)
- A Bayesian estimator of the proportion of non-heterosexuals aged $\geq$20 years was developed and compared with the method of moment (ME) and ML estimators
- RR data were collected by the Center for Survey Research, Academia Sinica in Taiwan
- A questionnaire containing 60 items was administered in face-to-face interviews
- One of the items, concerning sexual identity, was surveyed by means of Christofides' RR method

Christofides RR survey, 350 individuals

Respondents were posed the sensitive question

Q: Which number (X) refers to your sexual identity?

Number **0**: heterosexual

Number **6**: non-heterosexual (homosexual or bisexual)

and were asked

- not to reveal to the interviewer the number corresponding to Q and to bear it in mind
- to draw a card from a deck of 40 playing cards, where each card is marked with a number (Y) from 1 to 5 in the proportions $(0.2, 0.1, 0.2, 0.4, 0.1)$
- to release the absolute difference $D = |X - Y|$

Estimation for the proportion (θ) of Taiwanese non-heterosexual population

- Around the world, the most common estimates of the non-heterosexual population range from 1% to 10%, although the population under study is not always univocally identified and data are collected with different interviewing techniques
- The most recent data on sexual orientation released by the 2012 Taiwan Social Change Survey inform us that only **2.2%** of Taiwanese declare themselves to be non-heterosexual (0.3% homosexual and 1.9% bisexual)
- We have obtained Bayes estimates under three Beta prior distributions for θ: Beta (1,1), Beta (1,7) and Beta (3,40)

Estimation for the proportion (θ) of Taiwanese non-heterosexual population

Estimator	Mean	SD	95% CI	L	Mean	SD	95% CI	L
All sample, $n = 350$								
ME	0.157	0.348	[-0.525, 0.839]	1.364				
MLE	0.172	0.054	[0.067, 0.277]	0.210				
BE (1,1)	0.174	0.052	[0.073, 0.277]	0.204				
BE (1,7)	0.156	0.051	[0.061, 0.259]	0.199				
BE (3,40)	0.112	0.037	[0.047, 0.188]	0.141				
Gender		*Male,* $n = 172$ (49.1%)				*Female,* $n = 178$ (50.9%)		
ME	0.413	0.497	[-0.561, 1.387]	1.948	-0.090	0.487	[-1.044, 0.864]	1.908
MLE	0.084	0.073	[-0.058, 0.226]	0.285	0.264	0.078	[0.111, 0.417]	0.306
BE (1,1)	0.105	0.062	[0.009, 0.237]	0.228	0.269	0.079	[0.113, 0.427]	0.314
BE (1,7)	0.082	0.055	[0.005, 0.213]	0.209	0.222	0.075	[0.082, 0.370]	0.288
BE (3,40)	0.069	0.033	[0.017, 0.142]	0.125	0.123	0.049	[0.039, 0.228]	0.188

- Unacceptable and inaccurate ME
- Bayes estimates are more recommended than MLE
- Our study is expected to provide a more reliable and insightful analysis of the prevalence of non-heterosexuality in Taiwan, ranging from 11.2% to 17.4%
- The study has shown that women, young people aged 20-34, employed people, and people with senior high school education are those more affected by homosexual or bisexual identity

Thank you for your attention

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

pierfrancesco.perri@unical.it

References

- Chaudhuri A., Christofides T.C. (2013). *Indirect Questioning in Sample Surveys.* Springer-Verlag Berlin Heidelberg
- Christofides T.C. (2003). A generalized randomized response technique. *Metrika*, 57, 195-200
- Istat (2014a). Induced abortion, Year 2012. Available at http://www.istat.it/en/archive/141810
- Hsieh S.-H., Perri P.F. (2018). A logistic regression extension for the randomized response simple and crossed models: theoretical results and empirical evidence. Submitted
- Hsieh S.-H., Perri P.F. (2018). Estimating non-heterosexual identity in Taiwan using Christofides' randomized response method: some evidence from a real study. Submitted
- 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
- Perri P.F., Pelle E., Stranges M. (2016). Estimating induced abortion and foreign irregular presence using the randomized response crossed model. *Social Indicators Research*, 129, 601-618
- Perri P.F., Pelle E. (2018). Eliciting sensitive data via the randomized response model: some evidence from a study on the illicit cannabis use and its legalization. *Work in progress*
- Perri P.F., Rueda M., Cobo B. (2018a) Multiple sensitive estimation and optimal sample size allocation in the item sum technique. *Biometrical Journal*, 60, 155-163
- Perri P.F., Cobo B., Rueda M. (2018b). A mixed-mode sensitive research on cannabis use and sexual addiction: improving self-reporting by means of indirect questioning techniques. *Quality & Quantity*, 52, 1593-1611
- Rueda M., Perri P.F., Cobo B.R. (2018). Advances in estimation by the item sum technique using auxiliary information in complex surveys. *AStA-Advances in Statistical Analysis*, 102, 455-478
- Trappmann M., Krumpal I., Kirchner A., Jann B. (2011). Social desirability bias in survey measures of illicit work. In: 2nd International Conference of the German Priority Programme on Survey Methodology (PPSM) on Advancing Survey Methods. Bremen, November 17-18, 2011.
- 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