Assessing the quality of non-probability online panels in Italy.

Emanuela Sala and Chiara Respi
Università di Milano Bicocca

1. Motivations

Raising costs of data collection together with pressures for a timely delivery of research findings are amongst the main causes of the increasing use of web surveys in social, economic and health research (Schonlau and Couper 2017). However, when used to sample the general population, web surveys may pose serious methodological challenges, as there are currently no sampling frames available to survey the general population (Schonlau and Couper 2017). To overcome some of these limitations, researchers have started to collect survey data using non-probability online panels, i.e., panels of volunteer respondents who take part in surveys often in exchange of money. However, such panels pose a number of methodological issues; with few exceptions, research has consistently shown that they are not representative of the Internet or the general population they aim to represent and that estimates derived from the analysis of survey data collected using non-probability online surveys may be biased. We currently know very little on the processes that are driving the lack of representativeness of non-probability online panels, although a better understanding of these processes may lead to the collection of good quality survey data.

2. Research aims and contribution to the current knowledge

The overall aim of this paper is to investigate the impact of two sources of non-sampling error, i. e., undercoverage and nonresponse, on the representativeness of the Italian non-probability online panel *Opinione.net*, focussing, in particular, on nonresponse occurring at the following stages of the life of the panel: i) the recruitment stage, ii) the joining and profiling stage, and iii) the specific study stage. We analyse a unique set of data that includes information on all registered panellists. Being the first study that focuses systematically on selectivity introduced at different stages of the life of the panel, our research is an important contribution towards the improvement of data quality in non-probability online panels. This work contributes to enhance the current knowledge on this topic in a number of ways, being the first study that focuses on Italy and that critically discusses the very definition of Internet population, providing a new conceptualisation and implementation of this concept.

3. Data and methods

To pursue our aims, we use a diverse set of data, i. e., the 2015 Multipurpose Survey - Aspects of Everyday Living, considered as gold standard in our analysis, and two datasets from the non-probability online panel *Opinione.net*, i. e,. data on all registered panellists and data from the Italians' Living Conditions (ILC) survey, a study that was conducted on a quota sample of the *Opinione.net* panel members. We also use a diverse set of methods. We developed a new way to conceptualise and operationalise the concept of Internet population and computed and compared a number of data quality metrics (e. g. the largest absolute error), based on the results of the bivariate analysis. In our analysis, we focus on the following six socio-demographic variables, available in all three datasets: sex, age, marital status, education, occupation, and geographic area of residence.

4. Results and implications

A number of interesting findings stand out from our analysis. First, the ILC (responding) sample is not a representative sample of the Internet and, especially of the general population, the latter being characterised by a larger selection bias. However, weighting is very effective in reducing the magnitude of the bias, especially in the case of the general population. After weighting, the ILC sample is more representative of the general population than of the Internet population. This is good news. Although some bias remains even after weighting, in some cases non-probability online panels may be used to draw inference on the general population. However, more research is needed to further confirm these preliminary findings. Second, our study has also shown that the panel *Opinione.net* is not a representative sample of the Internet and, especially, of the general population, the latter being characterised by a larger bias. Third, it has also been documented that there are no major differences in the sample composition of the ILC (responding) sample and the ILC original sample. Again, this is a reassuring finding, confirming that nonresponse at the study specific stage is completely at random.

5. Limitations of the study

Our study has one main limitation. Our work is based on the analysis of one single non-probability online panel. We cannot exclude that extending the analysis to other panels may lead to different results. Researching the quality of non-probability online panels poses specific practical issues, because of the lack of (available) data on the panellists and on non-response at the study specific stage. The contribution of the market research industry is therefore key to a better understanding of the quality of the data collected using non-probability online panels.