María del Mar Rueda (University of Granada, Spain)
Beatriz Cobo (University of Granada)

Estimating sensitive behaviors using free software: the RRTCS package

Topic 2 – Learning more from what we already know

Keywords: randomized response techniques, R package, complex surveys, confidentiality, social desirability.

Introduction

In research, we very often gather information relating to highly sensitive issues. In these situations using the direct method of interview, the respondents provide often untrue response or even refuse to respond because of the social stigma and or fear. Such systematic response errors lead to social-desirability bias in prevalence estimates of the sensitive behaviors of interest, underestimating socially undesirable activities.

To overcome these problems, methods such as the randomized response technique (RRT) may be used to collect more reliable data, protect respondent's confidentiality and avoid unacceptable rate of nonresponse. In the RRT, respondents use a randomization device to generate a probabilistic relationship between their answers and the true values of the sensitive characteristic.

The RRT has been applied in surveys covering a variety of sensitive topics like racism, drug use, abortion, delinquency, AIDS or academic cheating. RR technique was originated by Warner in 1965 (Warner, 1965). Warner developed a data collection procedure, the randomized response (RR) technique that allows researchers to obtain sensitive information while guaranteeing privacy to respondents.

This method encourages greater cooperation from respondents and reduces their motivation to falsely report their attitudes.

Methods / Problem statement

Usually, RR methods are developed assuming that sample is obtained using simple random sampling. Most of the surveys in practice are complex surveys: involving stratification, clustering and unequal probability of selection of sample. Data from complex survey designs require special consideration with regard to estimation for finite population parameters and corresponding variance estimation procedures.

Standard software packages for complex surveys cannot be used directly when the sample is obtained from randomized response techniques. The analyses with standard statistical software, with certain modifications in the randomized variables, can yield correct point estimates of population parameters but still yield incorrect results for estimated standard errors. Recently some authors have developed R-packages for estimation with randomized response surveys.

The RRreg package (Heck and Moshagen, 2014), the rr package (Blair et al. 2015). The methods implemented in these packages are used under the assumption on simple random sampling and do not explore various theoretical and practical issues that may arise when adopting different survey sampling methods. To the best of our knowledge, there is no free software incorporating estimation procedures for handling randomized response data obtained from complex surveys.

Results / Proposed solution

RRTCS (Rueda et al. 2015) is a new R package to calculate point and interval estimation for linear parameters with data obtained from randomized response surveys. The package works with a wide range of sampling designs including simple random sampling with and without replacement (SRSWR and SRSWOR), stratified sampling, cluster sampling, unequal probabilities sampling and combination of them. It consists of twenty one main functions each of them implementing one of these RR procedures for complex surveys; the package also contains these functions considering domain estimation.

The package includes additional functions, one of them is IndirectEstimation, which calculates ratio, regression and difference estimators and the ResamplingVariance function, which provides estimates variance of the randomized response estimators using some resampling methods. Finally, the package includes datasets which contain observations from different surveys conducted in real and simulated populations using different randomized response techniques.

Conclusions

In his ground breaking paper in 1965, Warner proposed a very interesting idea of how to deal with evasive answer bias, especially when it comes to controversial survey questions. There have been many reports that RR provides more accurate estimates of the prevalence of socially undesirable behavior than does asking the sensitive question directly.

Numerous empirical studies have shown that RR obtains higher estimates of sensitive characteristics than are produced by direct questioning (Lara et al., 2006; van der Heijden et al., 2000). However, using RR incurs extra costs, and the advantage of using RR, i.e., the greater accuracy of the population estimates obtained, will only outweigh these extra costs if the estimates are substantially better than those derived from straightforward question-and-answer designs (Lensvelt-Mulders et al. 2005).

While immense methodological progress has been made in RR, there are only a handful of published studies that use the RR method to answer substantive question. Currently, for the estimation procedures there are many programs and programming languages for working with complex surveys, but there are few that have implemented modules to work with randomized response. Very recently there have appeared free programs that allow performing complex statistical analyses with data obtained by different randomized response techniques, like R-package RRreg, rr and RRTCS.