RANDOMISED RESPONSE TECHNIQUES FOR MEAN OF A QUANTITATIVE VARIABLE – THE APPLICATION TO CZECH WAGE DATA

Ondřej Vozár – Luboš Marek

Abstract

In surveys the respondents has been increasingly aware of their privacy, which leads to increasing non-response of even refusal to respond in all societies in the world. One ways to address this issue is use of randomized response techniques. In this paper, we focus on quite unexplored case of continuous quantitative variable with broad span of values (income). The short review of most common methods is done. The goal is to compare these methods with recent method by Antoch, Vozár and Mola. Because an individual income is perceived to be sensitive variable, the performance of the methods and choice of tuning parameter will be illustrated on wage distribution of the Czech Republic compiled by TREXIMA on behalf of national ministry of labor. The setting of the simulation study is the same as in Antoch et al. (2022) to satisfy comparability with their paper. Trade-off between respondent privacy and accuracy is shortly discussed.

Key words: randomized response techniques, Horvitz-Thompson estimator, survey sampling, population mean, wage distribution

JEL Code: C83, J30

Introduction

Decreasing response and growing concern about “invasion of privacy” by participants of the statistical surveys (respondents) have been observed around the world during last four decades. These issues have not been resolved regardless kind of survey: a) paper/e-mailed electronic questionnaire, b) face-to-face survey, c) interview by phone, c) internet surveys regardless additional procedures to reduce refusals to participate or more waves of callbacks.

Moreover, if a sensitive question is asked, quite high proportion of respondents refuses to answer or provide false/biased answers. Thus, standard techniques like model-based imputation (Särndal and Lundström, 2005) and reweighing (Brick, 2003) cannot resolve this issue. These obstacles can be partly mitigated by use of randomized response techniques

(RRTs). State of the art of RRTs is presented for example in comprehensive monograph Chaudhuri (2017). The first RRTs were proposed for estimation of proportion of sensitive variables. The first method of estimation of total/mean of sensitive quantitative variable was proposed by Eriksson (1973).

Aside two estimators proposed by Antoch et al. (2022) we present more common randomized response techniques for estimating population mean of a quantitative variable without use of any auxiliary information. The performance of these methods is studied by simulation study using data simulated from wage distribution of the Czech Republic in 2014 to extend simulation study in Antoch et al. (2022). Also choice of the tuning parameters for all the methods will be discussed very shortly.

1 Randomized Response Techniques for Population Mean

In survey sampling, the main goal is to estimate different characteristics of a finite population $U = \{1,2,\ldots,N\}$ of N unambiguously identified objects. For a sensititive quantatitative variable $\boldsymbol{Y}$ the objective is to estimate its population total $t_Y = \sum_{i\in U} Y_i$ or population mean $\bar{t}_Y = t_Y/N$. To achieve the goal, a random sample s of fixed sample size n is selected with probability $p(s)$. Using probabilities $\pi_i, (\pi_i = \sum_{s\ni i} p(s))$ of selection of ith unit of the population U, population mean is mostly estimated by linear unbiased Horvitz-Thompson estimator $\bar{t}_{Y,HT} = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i}$. If it is impossible to ask for values of variable $\boldsymbol{Y}$ directly, survey statisticians try to obtain at least randomized variable $\boldsymbol{Z}$ correlated to $\boldsymbol{Y}$. The randomization of responses is for each unit selected always carried out independently on the sampling procedure$p(s)$. Randomized response $\boldsymbol{Z}$ is further transformed to random variable $\boldsymbol{R}$, such as: $E_q(R_i) = Y_i$, $Var_q(R_i) = \phi_i$, for all $i \in U$, $Cov_q\left(R_i, R_j\right) = 0$, for all $i \neq j$ $i,j \in U$. E_q, Var_q and Cov_q are mean, variance and covariance with respect to probability distribution $q(r|s)$ of randomization of response of a selected sample s. Finally, population mean is estimated by unbiased Horvitz-Thompson estimator using R_i instead of Y_i as $\bar{t}_Y^R = \frac{1}{N}\sum_{i\in s}\frac{R_i}{\pi_i}$. Upper subscript R denotes the used randomised response technique. The RRTs applied later to Czech wage data are presented in subsequent subchapters.

1.1 Method of Eriksson (1973)

Eriksson (1973) proposed, that each respondent randomly selects a card from a deck of L cards with a numbers $b_1, b_2, \ldots, b_L$. The select card is unknown to an interviewer and ith respondent reports transformed value $b_i Y_i$ instead of original value Y_i. Size and values on deck of cards must be designed, that multiplicative transformation masks well values of sensitive variable $\boldsymbol{Y}$. More formally, it holds for selected units $Z_i^E = Y_i S_i^E, i \in s$, where S_i^E are independent identically distributed "scramble" random variables with $E_q(S^E) \neq 0$, $Var_q(S^E) \neq 0$. Then transformed randomised response is $R_i^E = Z_i^E / E_q(S^E)$ and population mean is estimated as $\bar{t}_Y^E = \frac{1}{N} \sum_{i \in s} \frac{R_i^E}{\pi_i}$.

1.2 Method of Chaudhuri (1987)

Chaudhuri (1987) modified proposal of Eriksson (1973). Each respondent randomly selects a card from two decks. The first deck consists of L cards with a numbers $b_1, b_2, ..., b_L$; the second deck consists of K cards with a numbers $c_1, c_2, ..., c_K$. Both selected cards are unknown to an interviewer and ith respondent reports transformed value $b_i Y_i + c_i$ instead of original value Y_i. It holds for selected units $Z_i^{CH} = Y_i S_{1,i}^{CH} + S_{2,i}^{CH}, i \in s$, where $S_{1,i}^{CH}$, $S_{2,i}^{CH}$ are independent identically distributed „scramble" random variables with finite, non-zero means and variances. Transformed randomized response is $R_i^{CH} = (Z_i^{CH} - E_q(S_2^{CH})) / E_q(S_1^{CH})$. Population mean is estimated as $\bar{t}_Y^{CH} = \frac{1}{N} \sum_{i \in s} \frac{R_i^{CH}}{\pi_i}$.

1.3 Method of Bar-Lev et al. (2004)

Bar-Lev, Bobovitch and Boukai (2004) modified proposal of Eriksson (1973) in following manner. With chosen probability p unkonwn both to respondent and interviewer, each respondent report its true value of sensitive variable Y_1. With probability $1 - p$ each respondent randomly selects a card from a deck of L cards with a numbers $b_1, b_2, ..., b_L$. The select card is unknown to an interviewer and ith respondent reports transformed value $b_i Y_i$ instead of original value Y_i. It holds for selected units ($i \in s$):

$$Z_i^{BBB} = \begin{cases} Y_i S_i^{BBB}, \text{ with probability } p \\ Y_i, \text{ with probability 1-}p \end{cases}, \tag{1}$$

where S_i^{BBB} are independent identically distributed „scramble" random variables with $E_q(S^{BBB}) \neq 0$, $Var_q(S^{BBB}) \neq 0$. If $p + (1 - p)E_q(S^{BBB}) \neq 0$, then transformed randomized response is $R_i^{BBB} = Z_i^{BBB} / E_q(S^{BBB})$. The population mean is unbiasedly estimated as $\bar{t}_Y^{BBB} = \frac{1}{N} \sum_{i \in s} \frac{R_i^{BBB}}{\pi_i}$. Trade-off between accuracy of estimates and protection of respondent privacy depending on probability p. The main concern is respondent privacy, therefore low values like $p = 0.1$ or $p = 0.2$ are used.

1.4 Methods of Antoch et al. (2022)

By our best knowledge, all randomized response techniques require respondent to provide transformed random variables. The calculation to provide such a transformed response may lead to refusals or errors. The requirement to provide transformed values of sensitive variable may also cause doubts in protection of respondent privacy. Proposal of Antoch et al. (2022) tries to overcome these obstacles. They assume that sensitive random variable $\boldsymbol{Y}$ is a positive, bounded ($0 < m \leq \boldsymbol{Y} \leq M$) and its bounds m, M are known.

For each respondent, pseudorandom number $\boldsymbol{U}$ on interval $[m, M]$ is independently generated. Its value is uknown for an interviewer. Respondent is only asked simple question "Is your income higher than value $\boldsymbol{U}$?". The response of ith respondent Z_i^A follows alternative distribution with parameter $\pi_i = (Y_i - m)/(M - m)$. Transformed randomized response is $R_i^A = m + (M - m)Z_i^A$; population mean is estimated as $\bar{t}_Y^A = \frac{1}{N}\sum_{i \in s}\frac{R_i^A}{\pi_i}$.

Antoch et al. (2022) also studied the case, if values of pseudorandom number $\boldsymbol{U}$ are known to interviewer, i.e. she/know also the question asked. With additional jeopardy of privacy accuracy of estimators can be improved a lot with use of pseudorandom numbers U_i.Response of ith respondents is modified as follows

$$Z_{i,\alpha}^A = \begin{cases} -1 + \alpha + 2\alpha\dfrac{U_i - m}{M - m}, & U_i < Y_i \\ \alpha + 2\alpha\dfrac{U_i - m}{M - m}, & \text{otherwise} \end{cases}, 0 \leq \alpha < 1, \qquad (2)$$

where α is a tuning parameter, apriori set by an interviewer and uknown to a respondent. Antoch et al. (2022) found optimal value of α with respect to variance of the estimator for simple random sampling without replacement. Optimal value of tuning parameter α depedends on value of population mean and population variance of a sensitive variable $\boldsymbol{Y}$. Numerical experiments in Antoch et al. (2022) shows that $\alpha = 0.75$ works quite well, if any prior information is available. Transformed randomized response is $R_{i,\alpha}^A = m + (M - m)Z_{i,\alpha}^A$; population mean is estimated as $\bar{t}_Y^{A,\alpha} = \frac{1}{N}\sum_{i \in s}\frac{R_{i,\alpha}^A}{\pi_i}$. Note a drawback of this proposal, that in rare cases some values of $R_{i,\alpha}^A$ can be negative, but it serves to estimate total of non-negative variable.

2 Application to Czech Wage Distribution

Income and wealth are recognized in many countries as private and sensitive information. Therefore, respondents are often prone to refuse to answer or to provide very biased answers. This particularly happens if their wealth or income is both low and high.

Therefore, we study performance of RRTs discussed above on Czech wage data of the Average Earnings Information System (ISPV) of the Ministry of Labor and Social Affairs of the Czech Republic for years 2014 to extend simulation study of Antoch et al. (2022).

Because of respondent privacy, no anonymized microdata files with the whole population or sample are provided to researchers, only frequency data with 100 CZK bin widths. Vrabec & Marek (2016) recommended to model wage distributions in the Czech Republic using a three-parameter log-logistic distribution with the density

$$f(y,\tau,\sigma,\delta) = \frac{\tau}{\sigma}\left(\frac{y-\delta}{\sigma}\right)^{\tau-1}\left(1+\left(\frac{y-\delta}{\sigma}\right)^{\tau}\right)^{-2}, y \geq \delta > 0, \tau > 0, \sigma > 0,$$

where $\tau > 0$ is a shape parameter, $\sigma > 0$ is a scale parameter and δ is a location parameter. Vrabec & Marek (2016) also estimates of the parameters of the distribution for the data of 2nd quarter 2014 as $\hat{\tau} = 4.0379$, $\hat{\sigma} = 21\ 687$, $\hat{\delta} = 250$. The parameters are estimated using over $2{,}1\ \times\ 10^3$ observations, the estimated mean wage is 24290 CZK.

This is why population U is generated using from model parametric wage distribution (log-logistic) using package *flexsurv*, see Jackson (2016). We run simulation study similarly as in Antoch et al. (2022); that 200 replications populations with size $N = 400$ and $N = 200$ are simulated from log-logistic model, with parameters estimates given above. Let us note, that population sizes $N = 400$ and $N = 200$ are common sizes of surveyed community (group of students, size of village or small community, etc.).

Tab. 1: Choice of tuning parameters and scramble variables

Method	**Tuning parameters**
Eriksson (1973)	$S^E \sim Unif[0.25,2]$
Chaudhuri (1987)	$S_1^{CH} \sim Unif[0.25,2], S_2^{CH} \sim Unif[-10000,10000],$
Bar-Lev et al. (2004)	$S^{BBB} \sim Unif[0.25,2], p = 0.1$
Antoch et al. (2022) – basic proposal	$U \sim Unif[8000,60000]$
Antoch et al. (2022) – with use of ***U***	$U \sim Unif[8000,60000], \alpha = 0.75$

Source: The authors.

From each simulated population we draw 200 random samples without replacement of sizes $n = 50$, $n = 20$. The sample size are standard sample sizes in official statistics (both household and business surveys). All calculations and simulations are done by statistical freeware R, version 4.2.0, see R Core Team (2021). Choice of tuning parameters and scramble variables (see Table 1) takes into account nature of the wage data (many values on broad range). We used continuous uniformly distributed scramble variables, because decks of cards would be very large to mask monthly wage. Parameters for methods by Antoch et al. (2022) are used. The bounds m, M for pseudorandom number $\boldsymbol{U}$ were set with regards, which values of salaries can be perceived high. While 8 000 CZK corresponds to the 0.01 sample quantile and 60 000 CZK corresponds to the 0.97 sample quantile, which justifies this choice of m and M. Because of no prior information, fixed value of tuning parameter is set to $\alpha = 0.75$.

2.1 Main results

Numerical results of simulations and behavior of RTTs discussed in the first chapter are presented in Table 2 and Figure 1.

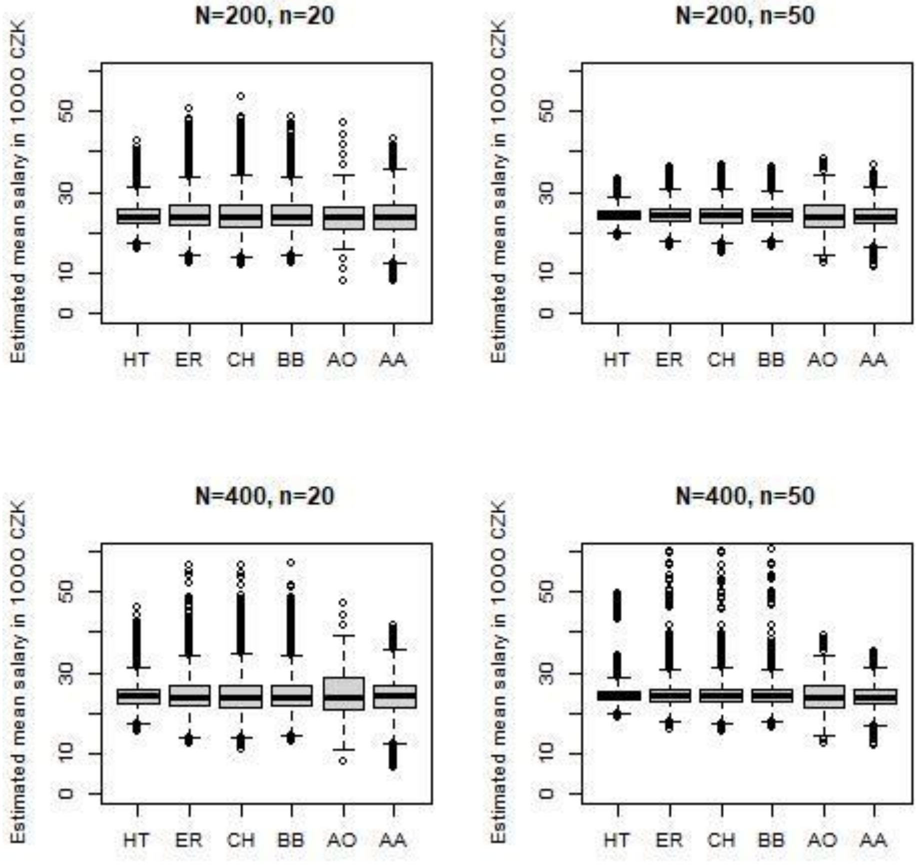
Tab. 2: Numerical results of simulations – Czech wage data, 2nd quarter 2014

Estimator		$N = 200$		$N = 400$	
		$n = 20$	$n = 50$	$n = 20$	$n = 50$
$\bar{t}_{Y,HT}$	mean	24.291	23.313	24.239	24.261
	sd	2.721	1.726	2.739	1.730
$\bar{t}_Y^E$	mean	24.285	24.334	24.228	24.272
	sd	3.847	2.448	3.863	2.437
$\bar{t}_Y^{CH}$	mean	24.277	24.333	24.220	24.271
	sd	4.007	2.553	4.019	2.540
$\bar{t}_Y^{BBB}$	mean	24.299	24.229	24.238	24.270
	sd	3.795	2.401	3.795	2.395
$\bar{t}_Y^A$	mean	24.005	24.016	23.986	23.971
	sd	5.362	3.373	5.362	3.394
$\bar{t}_Y^{A,\alpha}$	mean	23.989	24.024	23.969	23.979
	sd	4.403	2.772	4.398	2.773

The mean estimated wages (in thousands of CZK) and corresponding standard deviations for different population and sample sizes. Means and standard deviations (sd) are averaged over 200 x 200 samples.
Source: The authors

Horvitz-Thompson mean estimator with full response is included in the study as benchmark. Use of RRTs instead of direct questioning causes decreasing precision for protection of privacy. The decrease in precision is in line with level of protection of respondents´ privacy by a given RRT.

Fig. 1: Behavior of assessed estimators for different population (N) and sample (n) sizes



HT: $\bar{t}_{Y,HT}$, ER: $\bar{t}_Y^E$, CH: $\bar{t}_Y^{CH}$, BB: $\bar{t}_Y^{BBB}$, AO: $\bar{t}_Y^A$, AA: $\bar{t}_Y^{A,\alpha}$.

Source: The authors

In comparison with direct questioning with full response, the sample standard deviation of estimates using RRTs is increased approximately by:

- 41% for method of Eriksson (1973),
- 47% for method of Chaudhuri (1987),
- 39% for method of Bar-Lev et al. (2004),
- 96% for original proposal of Antoch et al. (2022),
- 60% for proposal of Antoch et al. (2022) with use of pseudorandom numbers $\boldsymbol{U}$.

These accuracy is satisfactory, if you consider decrease in accuracy caused by high non-response like 50% and higher (experienced in statistical surveys) for direct questioning.

The higher protection of respondent privacy, the higher decrease in accuracy of estimates. The accuracy of RTTs by Eriksson (1973) and Bar-Lev et al. (2004) for reasonable values of probability p is practically the same. So proposal of Bar-Lev et al. (2004) is mostly generalization of much simpler and more trustworthy Eriksson (1973) method. Decrease in accuracy of much safer original proposal by Antoch et al. (2022) is quite high, but the modification with knowledge of question asked (with use of pseudorandom numbers $\boldsymbol{U}$) is quite competitive with Chaudhuri (1987) proposal. The increase in accuracy with use of pseudorandom numbers $\boldsymbol{U}$ is almost 20% in terms of standard error of estimated means.

All methods using scramble variables provide unbiased estimates of mean salary. Both methods of Antoch et al. (2022) provide slightly biased estimates (approx. smaller by 250 CZK), because more than 4% of data lies outside the interval $[8000,60000]$. As noted in Antoch et al. (2022) the proper choice of interval for pseudorandom numbers $\boldsymbol{U}$ is vital for its use. There is a trade-off between bias and variance.

All estimators using RTTs suffer from outliers (See Fig. 1), the presence of extremely high estimates improves with growing sample size. Outliers with very low values using method of Antoch et al. (2022) with use of pseudorandom numbers $\boldsymbol{U}$ are caused by presence of negative values of $R_{i,\alpha}^{A}$. Effects of setting negative values to zero on bias and variance have not been studied yet.

Conclusion

The paper provides an overview of most common randomized response techniques to estimate population mean and an assessment of their performance on real data – the Czech wage distribution. The performance of estimators depends on choice of scramble variables and tuning parameters. Their choice is shortly discussed in the paper. We are aware that the key issue to gain trust of the respondents. This is difficult, because all RRTs presented can be seen as cumbersome and demanding for respondents (transforming sensitive variables) and infringement of their privacy (they that interviewer can guess their answer using computers).

Acknowledgment

Work of both authors was supported by Institutional Support to Long-Term Conceptual Development of Research Organization, the Faculty of Informatics and Statistics of the University of Economics, Prague.

References

Antoch, J., Mola, F., & Vozár, O. (2022). New Randomized Response Technique for Estimating the Population Total of a Quantitative Variable. *Statistika: Statistics and Economy Journal,* 102(2), 205-227. https://doi.org/10.54694/stat.2022.11

Bar-Lev, S., K., Bobovitch, E., & Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60, 255-260. https://doi.org/10.1007/s001840300308

Brick, M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329–353.

Chaudhuri, A. (1987). Randomized response surveys of a finite population: A unified approach with quantitative data. *Statistical Planning and Inference*, 15, 157–165.

Chaudhuri, A. (2017). *Randomized Response and Indirect Questioning Techniques in Surveys.* New York: Chapman and Hall/CRC. ISBN 978-11-3811542-2.

Eriksson, S. (1973). A new model for randomized response. *International Statistical Review*, 41, 101–113.

Greenberg, B. (1971). Application of the randomized response technique in obtaining quantitative data. *J. American Statistical Association*, 66, 243–250.

Jackson, C. (2016). *flexsurv:* A platform for parametric modelling in R. *J. of Statistical Software*, 70, 1–33.

R CORE TEAM (2022). *R: A language and environment for statistical computing.* Austria, Vienna: R Foundation for Statistical Computing.

Särndal, C., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: J. Wiley and Sons. ISBN 978-0470-01133-1.

Tillé, Y. (2020). *Sampling and Estimation from Finite Populations.* New York: J. Wiley and Sons. ISBN 978-0470-68205-0.

Vrabec, M., & Marek, L. (2016). *Model of distribution of wages.* AMSE 2016, 19th Symp. Applications of Mathematics and Statistics in Economics, Banská Štiavnica, 378–396.

Contact

Ondřej Vozár
Prague University of Economics and Business, Faculty of Informatics and Statistics
Department of Statistics and Probability
W. Churchill Sq. 1938/3, 130 67 Prague, Czech Republic
vozo01@vse.cz

Luboš Marek
Prague University of Economics and Business, Faculty of Informatics and Statistics
Department of Statistics and Probability
W. Churchill Sq. 1938/3, 130 67 Prague, Czech Republic
marek@vse.cz