Application of Randomized Response Techniques Using Dichotomous Response for Mean Wage in Czechia and Slovakia

Ondřej Vozár[1] | *Prague University of Economics and Business, Prague, Czechia*
Luboš Marek[2] | *Prague University of Economics and Business, Prague, Czechia*

Received 18.8.2024, **Accepted** (reviewed) 9.9.2024, **Published** 14.3.2025

Abstract

Research of controversial topics (drug consumption, corruption) requires reliable estimates of population means of sensitive variables (spending on drugs, illegal sources of income). To avoid non-response and fabricated responses surveys using randomized response techniques (RRT) for quantitative variables are conducted. The paper focuses on surveys conducted regularly in time to study evolution of population mean of a sensitive variable. This topic has not been explored for RRT yet. In applications of RRT is critical a choice of its parameters. The goal is to find rules of thumb if mean of a sensitive variable change in time. We focus only on methods using dichotomous variable (Antoch et al., 2022), because they were designed for variables with many values (Vozár and Marek, 2023). The different scenarios are applied on the Czech and Slovak wage data from Average Earnings Information System in years 2017–2019 using prior information. The scenarios evaluated in extensive simulation study focusing both on mean wage and year-on-year growth.[3]

Keywords	***DOI***	***JEL code***
Randomized response techniques, dichotomous response, wage distribution, survey sampling, comparability over time, population mean	*https://doi.org/10.54694/stat.2024.42*	*C83, C10, E24*

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Churchill Sq. 4, 130 67 Prague 3, Czechia. E-mail: vozo01@vse.cz.
[2] Department of Statistics and Probability, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Churchill Sq. 4, 130 67 Prague 3, Czechia. E-mail: marek@vse.cz. ORCID: <https://orcid.org/0000-0003-4761-1936>.
[3] Article based on the *AMSE 2023* Conference contribution.

INTRODUCTION

The field randomized response techniques (RRT) for quantitative variables have been rapidly developing both in theory and application (e.g. Chaudhuri and Christfides, 2013; or Chaudhuri et al., 2016) since the first pioneering paper fifty years ago by Eriksson (1973). Vozár and Marek (2023) summarized three main approaches:

- Methods using scramble variables (Eriksson, 1973; Eichorn and Hayre, 1983), where instead of true values respondents provide linearly transformed values of sensitive variables depending on results of a random experiment.
- Methods using scramble variables using auxiliary variables known for the whole population strongly correlated to estimated sensitive variables (Diana and Perri, 2013).
- Methods using dichotomous response (Antoch et al., 2022), where respondent provides only dichotomous response ("Yes/No") instead of any numerical value related to value of studied sensitive variables.

They argument, that methods using dichotomous response (Antoch et al., 2022) are more fit for sensitive quantitative variables with broad range and numbers of its values (i.e. variables in financial units). They also provide more comfort to respondents (no need of numerical calculations like in methods using scramble variables) and protection of confidentiality of respondents' data. We also exclude techniques using auxiliary variables, because their use is not feasible in most of the real-life populations. The main objection is, that if auxiliary variable is strongly correlated with the sensitive variable, the auxiliary variable must be also sensitive. Therefore, such an auxiliary variable would be unavailable or available in very poor quality.

The rest of the paper is organized as follows. Section 1 summarizes basic notions of survey sampling of a finite population, principles of randomized response techniques, estimation of population mean in this setting, notations and methods using dichotomous responses by Antoch et al. (2022). Section 2 deals with presentation of wage data, including Average Earning Information Systems of Czechia and Slovakia, evolution of wage distribution in studied period of 2016–2019, evolution of wage distribution and choice of wage distribution model. In Section 3, different scenarios for parameter choice of RRT are proposed, setting of simulation study and its numerical results are discussed. The last section summarizes the main findings and conclusions of the paper.

1 EVALUATED METHODS OF DICHOTOMOUS RESPONSES

This section presents basic notion of survey sampling and randomized response techniques, brief review of RRT for population mean and studied methods using dichotomous responses by Antoch et al. (2022).

1.1 Basic notions of survey sampling and randomized response techniques

The goal of survey sampling is to estimate characteristics of a finite population $U = \{1,2,\ldots,N\}$ of N unique objects (units). For a quantitative variable Y it is its population total $t_Y = \sum_{i\in U} Y_i$ or population mean $\overline{t}_Y = t_Y / N$ mostly. To achieve that, a random sample s of fixed sample size n is selected with probability $p(s)$. Using probabilities $\pi_i, \left(\pi_i = \sum_{s\ni i} p\left(s\right)\right)$ of selection of i^{th} unit of the population U, unbiased Horvitz-Thompson estimator then estimates population mean:

$$\overline{t}_Y^{HT} = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i}, \tag{1}$$

where subscript HT refers to type for the estimator (1). Statistical properties of estimators and proofs are presented in Section 2.8 in Tillé (2006). If the surveyed variable is sensitive, respondents often refuse to answer or provide fabricated answers. Instead, interviewers try to obtain randomized variable Z correlated

to variable of interest Y. Randomization of responses is always carried out independently for each unit selected in sample s with probability $p(s)$. Randomized responses Z are then transformed to random variables R following standard model by Arnab (1994):

$$E_q(R_i) = Y_i, Var_q(R_i) = \phi_i \text{ for all } i \in U, Cov_q(R_i, R_j) = 0, \text{ if } i \neq j \; i, j \in U, \tag{2}$$

where E_q, Var_q and Cov_q denote 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 transformed randomized responses R_i instead of values of sensitive variable Y_i:

$$\overline{t}_Y^{HT,R} = \frac{1}{N}\sum_{i \in s}\frac{R_i}{\pi_i}, \tag{3}$$

where upper subscript R denotes the used randomized response technique.

1.2 Methods using dichotomous responses

Standard methods using scramble variables have several drawbacks (Antoch et al., 2022; Vozár and Marek, 2023):

- Missing practical guidelines for designing scramble variable.
- Calculations required from respondents can be too demanding or misleading for respondents (they can lead to severe errors or non-response).
- Method can be less trustworthy for respondents, because they can feel that interviewer can guess somewhat the sensitive value. In addition, if with knowledge of the values of scramble variable, true value of sensitive variable can be directly calculated.

To resolve the drawback, Antoch et al. (2022) proposed completely different approach. Assuming that the surveyed sensitive variable Y is non-negative and bounded from above ($0 < m \leq Y \leq M$) and both bounds m, M of the variable Y are known. Each respondent draws (independently of the others), a random number U from the uniform distribution on interval (m,M). The interviewer does not know this value. Finally, the respondent answers only a simple question: "Is the value of Y greater than U?" For example: "Is your monthly income greater than U?"

Note, that even if an interviewer knows the value of random number U, he cannot guess the true value of U accurately (unless $Y=U=M$). Therefore, they proposed more accurate estimator using the values of random numbers U. Vozár (2023) derived unbiased variance estimators using plug-in technique of Arnab (1994) by assuming knowledge of random numbers U. Without use of random numbers U approximate confidence intervals can estimated by using computer-intensive methods like bootstrap.

1.2.1 Original method of Antoch et al. (2022)

Randomized response of i^{th} respondent follows alternative distribution with parameter $\frac{y_i - m}{M - m}$:

$$Z_{i,\alpha,(m,M)} = \begin{cases} 1 - \alpha + 2\alpha\dfrac{U_i - m}{M - m} & \text{with probability } \dfrac{y_i - m}{M - m}, \text{ if } U_i < y_i, \\ -\alpha + 2\alpha\dfrac{U_i - m}{M - m} & \text{with probability } 1 - \dfrac{y_i - m}{M - m}, \text{ if } U_i \geq y_i, \end{cases} \tag{4}$$

Transformed randomized response is then given as:

$$R_{i,(m,M)} = m + (M - m) Z_{i,(m,M)}. \tag{5}$$

Unbiased population mean estimator of Horvitz-Thompson type is then:

$$\overline{t}_{Y,(m,M)}^{HT,R} = \frac{1}{N}\sum_{i\in s}\frac{R_{i,(m,M)}}{\pi_i}. \tag{6}$$

1.2.2 Method of Antoch et al. (2022) using values of random numbers

Randomized response of i^{th} respondent incorporates information on random number in the following manner:

$$Z_{i,\alpha,(m,M)} = \begin{cases} 1-\alpha+2\alpha\dfrac{U_i-m}{M-m} & \text{with probability} \quad \dfrac{y_i-m}{M-m}, \text{ if } U_i < y_i, \\ -\alpha+2\alpha\dfrac{U_i-m}{M-m} & \text{with probability } 1-\dfrac{y_i-m}{M-m}, \text{ if } U_i \geq y_i, \end{cases} \tag{7}$$

where α is a tuning parameter. Its value is a priori set by the interviewer, is fixed and unknown to the respondent. Antoch et al. (2022) derived its optimal value minimizing variance for case of sample with constant selection probabilities π_i. Vozár (2024) found out in extensive simulation study that values $\alpha = 0.5$ or $\alpha = 0.7$ provided good results for broad class of distributions of sensitive variables.

Transformed randomized response is then equal to:

$$R_{i,\alpha,(m,M)} = (M-m)Z_{i,\alpha,(m,M)} + m. \tag{8}$$

Unbiased population mean estimator of Horvitz-Thompson type is then:

$$\overline{t}_{Y,\alpha,(mM)}^{HT,R} = \frac{1}{N}\sum_{i\in s}\frac{R_{i,\alpha,(m,M)}}{\pi_i}. \tag{9}$$

2 CZECH AND SLOVAK WAGE DATA AND ITS STATISTICAL DISTRIBUTION

In this section the studied Czech and Slovak Wage data and statistical model of wage distribution are presented. Since no anonymized microdata on wages are available, we simulate the corresponding populations from the estimated wage distribution. First, statistical surveys Average Earnings Information Systems (ISPV) in both countries are shortly presented. Then, the evolution of Czech and Slovak wage distributions in years 2016–2019 is summarized. The last subsection discusses the methods of modelling Czech and Slovak wage data from Average Earnings Information Systems and provides estimated wage distributions for studied data. To achieve comparability, Slovak wage data were converted to Czech crowns.

2.1 Average Earnings Information System (ISPV) in Czechia and Slovakia

The wage data used come from corporate statistical surveys conducted within the Czech and Slovak statistical services, always on behalf the national Ministry of Labour and Social Affairs. Our study will focus to the so-called wage sphere, i.e. mainly wages in the private sector, while the data cover the population of employers with ten or more employees.

The sampling plans in both countries aim to cover the largest possible volume of wages paid with the lowest possible range of company selections and, at the same time, to provide representative results for the size groups of companies in terms of the number of employees. Data for size groups are important for economic policymaking and for other stakeholders:

- Census of large enterprisers with 250 and more employees.
- Survey sampling for the size group of medium-sized enterprises with 50 to 249 employees and small enterprises with 10 to 49 employees.

- Once every two years, a supplementary survey is carried out on the smallest enterprises with 1 to 9 employees, however, data for these enterprises are not included in the data from which the wage distributions studied in this chapter are estimated.

Although a small number of employers are selected, the ISPV data cover a substantial part of employees in the wage sphere of the Czech and Slovak economies (over 2.20 million of employees in Czechia, over 1.05 million of employees in Slovakia). In this study, we will deal with the average gross monthly wage in the second quarter of the given year. The second quarter is chosen as the reference period when studying wage distributions, as in this period wages are not affected by bonuses paid and transfers of non-entitlement components of wages to other quarters due to a change in legislation (in the case of a change in taxation, these components are always paid in the period when it is tax-advantageous).

2.2 Changes in the Czech and Slovak wage distribution in years 2016–2019

This chapter contains a summary of the development of wage distributions in the period under review, including the possible impact of the increase in the minimum wage on wage growth (see Table 1), which was significant in both countries in this period.

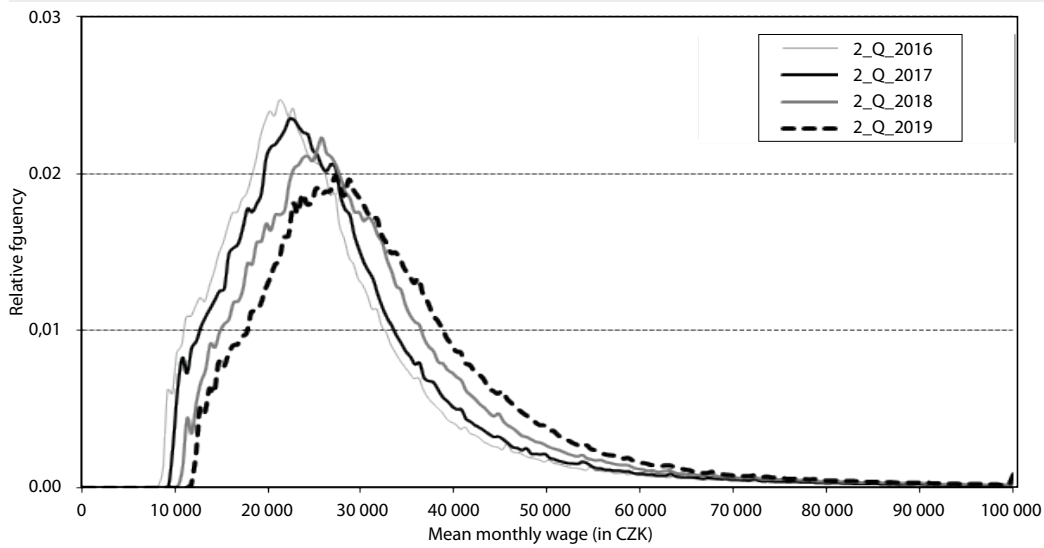
Table 1 Minimum wage in CZ and SK in years 2016–2019 (in CZK)

Country	Indicator	Year			
		2016	2017	2018	2019
CZ	Minimal wage (CZK)	9 900	11 000	12 200	13 350
SK		10 125	10 875	12 000	13 000

Source: Ministry of Labour and Social Affairs (CZ): <*www.mpsv.cz*>, Ministry of Labour, Social Affairs and Family (SK): <*www.employment.sk*>

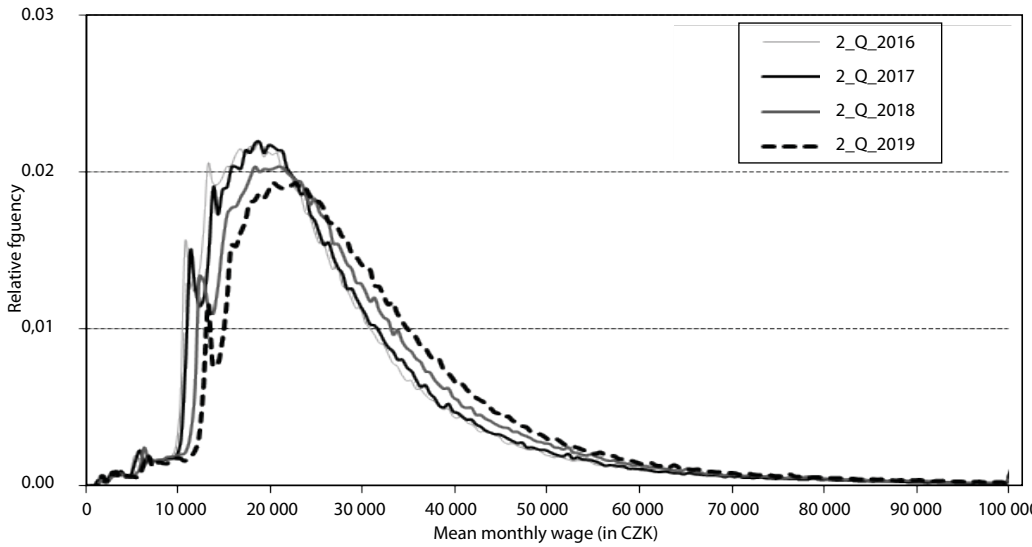
The wage distributions in both countries are strongly skewed from the right and indicate high wage equality (see the graphs in Figures 1 and 2).

Figure 1 Histogram of mean monthly wages in the second quarter wages in Czechia in years 2016–2019



Source: Own construction using interval data by company TREXIMA

Figure 2 Histogram of mean monthly wages in the second quarter wages in Slovakia in years 2016–2019



Source: Own construction using interval data by company TREXIMA

Wage levels and inequality are higher in Czechia than in Slovakia (see Table 2).

Table 2 Wage distribution in CZ and SK in years 2016–2019

Country	Mean monthly gross wage (CZK)	Period			
		2Q 2016	2Q 2017	2Q 2018	2Q 2019
CZ	1. decile	13 694	14 732	16 675	18 364
	1. quartile	18 141	19 302	21 465	23 448
	median	23 506	24 886	27 490	30 088
	3. quartile	30 710	32 361	35 625	39 080
	9. decile	41 779	44 082	47 888	52 099
SK	1. decile	12 778	13 256	14 398	15 830
	1. quartile	16 506	17 105	18 482	20 064
	median	22 493	23 079	24 901	26 828
	3. quartile	31 663	32 386	34 674	36 925
	9. decile	46 296	47 006	49 517	51 995

Source: Own construction from interval data by company TREXIMA

Both countries recorded high wage growth (most notably in 2018 and 2019), driven by minimum wage increases (see Table 1 and shift to the right in Figures 1 and 2). The increase was in lower wage bands and therefore there was a flattening and shift of wage distributions to the right (see Figures 1 and 2). The share of employees with an average monthly wage exceeding CZK 100 000 grew rapidly (over 3% in 2019 in both countries, see Figures 1 and 2).

2.3 Choice of wage distribution model

As anonymized microdata for scientific purposes from ISPV are not available (including that personal data covered by the GDPR), simulated data from the estimated parametric wage distribution from interval data must be used to study the proposed estimates.

Modelling of wage distributions in Czechia and Slovakia based on interval data from national ISPV statistical surveys has been the subject of a number of papers in which a suitable statistical methodology has been sought to find adequate models and changes associated with the transformation of both economies have been studied.

Marek (2010), Vrabec and Marek (2016) dealt with the choice of a suitable parametric model to capture the development of wage distributions. A three-parameter log-logistic distribution (see Formula (10)) has been shown as a suitable model, which will be used for simulations in this chapter due to its versatility for empirical data and computational simplicity. Bílková and Malá (2012), Bílková (2013) modelled wage distributions for individual industries using so-called L-estimates. Modelling of multivariate distributions (Malá, 2015) and finite mixtures of distributions (Malá, 2013; Marek and Vrabec, 2016) has proven to be a suitable methodology for modelling and analysing the development of wage distributions. Estimating the parameters of wage distributions by these models would be too complicated to generate the studied variables, so we chose a computationally simpler approach using a one-dimensional parametric model, namely a three-parametric log-logistic distribution with density:

$$f\left(y,\tau,\sigma,\delta\right)=\frac{\tau}{\sigma}\left(\frac{y-\delta}{\sigma}\right)^{\tau-1}\left(1+\left(\frac{y-\delta}{\sigma}\right)^{\tau}\right)^{-2},\ y\geq\delta>0,\tau>0,\sigma>0, \tag{10}$$

where $\tau > 0$ is a shape parameter, $\sigma > 0$ is a scale parameter and $\delta > 0$ je is a shift parameter.

The parameters of the three-parameter log-logistic distribution (10) were estimated by the moment method (see Table 3 for parameter estimates). All estimates were based on interval data in the range of CZK 0 to 100 000. Employees with an average monthly wage exceeding one hundred thousand. CZK, we have included in the last interval (CZK 99 501–100 000).

Table 3 Estimates of parameters of wage distribution in CZ and SK in 2016–2019

Country	Year	Parameter		
		$\hat{\tau}$	$\hat{\sigma}$	$\hat{\delta}$
CZ	2016	3.5	19 812	3 736
	2017	3.6	21 198	3 737
	2018	3.8	23 831	3 742
	2019	4.0	26 397	3 738
SK	2016	3.5	24 473	389
	2017	3.6	23 507	–12
	2018	3.7	25 378	–122
	2019	4.0	28 171	–980

Note: Model wage distribution is three-parametric log-logistic distribution (10), parameters estimated by moment methods form interval data by company TREXIMA.
Source: Own construction

3 SIMULATION STUDY

In the first subsection, we choose statistics for performance evaluation of estimates. In the second subsection, we propose the strategies for setting intervals (m, M) using prior information on minimal

wage. In the third section numerical simulation results are presented and discussed. Because the aim of the paper is to find a general rule of thumb, the study is restricted period of rapid wage growth in years 2016–2019. The economic shocks due to Covid–19 pandemics and corresponding measures on labour market would complicate finding this rule of thumb.

3.1 Statistics for simulation evaluation

Assume, that is given simulation M replications are carried out and statistics $S_i,\ i = 1, 2, \ldots, M$ (population mean) are in i^{th} replica estimated by $\hat{S}_i,\ i = 1, 2, \ldots, M$ (sample mean). The statistics below is used to evaluate estimates $\hat{S}_i,\ i = 1, 2, \ldots, M$.

To evaluate bias, mean percentage (MPE) is defined as:

$$MPE = \frac{1}{M}\sum_{i=1}^{M}\frac{\hat{S}_i - S_i}{S_i},\ S_i \neq 0,\ i = 1, 2, \ldots, M. \tag{11}$$

To evaluate variance of estimates, median absolute percentage error (MdAPE) is defined (Hyndman a Koehler, 2006) as:

$$MdAPE = med\left(\left|\frac{\hat{S}_1 - S_1}{S_1}\right|, \left|\frac{\hat{S}_2 - S_2}{S_2}\right|, \ldots \left|\frac{\hat{S}_M - S_M}{S_M}\right|\right),\ S_i \neq 0,\ i = 1, 2, \ldots, M. \tag{12}$$

3.2 Strategies for setting intervals (*m*, *M*)

The aim of the application is to estimate a sensitive variable – the average monthly salary in the second quarter in a time series, which is a real application in the practice of state statistics or statistical agencies. The strategy differs according to the rules for choosing the interval (m, M) for the generation of random numbers, the value of the parameter α for the method using the knowledge of the random number is set to $\alpha = 0.5$ and 0.35 using recommendation of simulation study by Vozár (2024). The following strategies are proposed:

- S1: fixed interval values (m, M) for the whole period 2016–2019. The bounds are chosen ad hoc according to how we perceive too low or too high a wage.
- S2: fixed upper limit of the interval (m, M) for the whole period 2016–2019, the lower limit is equal to the minimum wage valid in the second quarter of the given year.

Multiple intervals (m, M) are to be evaluated for each strategy:

Table 4 Intervals (*m, M*) for random numbers

Strategy		Country	Year			
			2016	2017	2018	2019
S1	low	CZ, SK	(10 000, 50 000)			
	medium		(10 000, 60 000)			
	high		(12 000, 70 000)			
S2	low	CZ	(9 900, 50 000)	(11 000, 50 000)	(12 200, 50 000)	(13 350, 50 000)
	medium		(9 900, 60 000)	(11 000, 60 000)	(12 200, 60 000)	(13 350, 60 000)
	high		(9 900, 70 000)	(11 000, 70 000)	(12 200, 70 000)	(13 350, 70 000)
	low	SK	(10 125, 50 000)	(10 875, 50 000)	(12 000, 50 000)	(13 000, 50 000)
	medium		(10 125, 60 000)	(10 875, 60 000)	(12 000, 60 000)	(13000, 60 000)
	high		(10 125, 70 000)	(10 875, 70 000)	(12 000, 70 000)	(13 000, 70 000)

Source: Own construction

No that there is a bias-variance trade-off of population mean estimates (Antoch et al., 2022). The smaller range of the interval (m, M), the lower variance and mostly higher bias and vice versa.

3.3 Simulation study and results

We focus on a single combination of sample range and population size, namely sample range $n = 1\,000$ and population size $N = 1\,000\,000$. All simulations were carried out by statistical freeware *R* (R Project, 2024). Wage data were generated by R package *flexsurv* (Jackson, 2016) following three-parametric log-logistic distribution (Formula 10) with parameters estimated by moment method (Table 3). The size of the sample is motivated, by the typical sample size for a national survey conducted by public opinion research agencies. We chose fixed population sizes for two reasons. The first reason is easier comparability of results in individual years, because this ensures the same size of the population and the same sampling ratio, namely one per mile. The second reason is the acceleration of simulations, and we have shown by numerical studies that for a sample ratio of one per mile, the results would not differ much from each other. We will limit ourselves to assessing the combination of methods and strategies for the choice of interval (m, M) (Table 4) from the point of view of impartiality and variability of estimates of average monthly wages. Effect of non-response is studied to provide benchmarks with direct questioning in this setting:

- Missing completely at random (MCAR), where 90 %, 80 % or 60 % respondents of sample answers.
- Missing not at random (MNAR), where 10 % of respondents with the highest respondents refuses to answer (we assume high values of wages as more sensitive)

We treat non-response by weighting, mean wage is estimated as sample mean of responses.

For wage data, the impact of systematic non-response is severe. The relative bias represents approximately one eighth of the average wage and an almost eleven-fold increase in variability compared to the 100% response in direct questioning (see Table 5).

Table 5 Effect non-response to bias and accuracy ($N = 1\,000\,000$, $n = 1\,000$)

Country	Year	100% response (MdAPE)	Non-response model				
			MCAR (MdAPE)			MNAR ($n_r = 0.9n$)	
			$n_r = 0.9n$	$n_r = 0.8n$	$n_r = 0.6n$	MdAPE	MPE
CZ	2016	1.14	1.20	1.27	1.46	12.60	–12.59
	2017	1.10	1.16	1.23	1.41	12.15	–12.15
	2018	1.03	1.08	1.15	1.33	11.40	–11.40
	2019	1.03	1.08	1.15	1.33	11.40	–11.10
SK	2016	1.29	1.36	1.44	1.67	14.34	–14.34
	2017	1.26	1.32	1.41	1.61	13.95	–13.95
	2018	1.23	1.29	1.37	1.56	13.55	–13.54
	2019	1.14	1.19	1.27	1.46	12,61	–12,60

Source: Own construction

The S1 strategy with fixed intervals *(m, M)* is not satisfactory due to the high wage growth in the period under review, because at the end of the period the estimates $\overline{t}_{Y,(m,M)}^{HT,R}$ would be very biased (see Table 6). The S2 strategy, with an annual update of the lower limit of the interval with the applicable minimum wage, eliminates this bias only slightly (see Table 8). Moreover, the bias (MPE) changes over time with rising wage levels, leading to a strongly biased estimate of annual average wage growth.

For wage data, heuristics motivating estimated with use random number (Antoch et al., 2022) works very well. This leads to a significant reduction in the underestimation of the average wage (see Tables 6 and 8), only 1.7–2.5%, regardless of the value of the parameter α. To balance the bias and variance of estimates, it is necessary to choose the highest possible upper limit, i.e. CZK 70 000. If we choose an upper limit of only CZK 50 000 (the ninth decile), then the lower limit must be much higher than the values we consider compensating for the bias caused by wages exceeding this upper limit.

Considering the variability of estimates, a combination of the "high" variant with the highest upper bound and an estimate using the knowledge of random numbers is appropriate (see Table 7 and 9). It is also advisable to set the parameter $\alpha = 0.5$, i.e. the contribution to the variance caused by randomized response is constant, it does not depend on the value of the sensitive variable y_i. Then the bias of the estimates (MPE) is the same over time, which allows for unbiased estimates of year-on-year growth rate. Due to the sharp increase in wages in the period under review, we are inclined to the S2 strategy with an update of the lower bound of the interval *(m, M)*.

Table 6 Mean percentage error (MPE) – strategy S1

Estimate	Strategy for *(m, M)*	Country	Year			
			2016	2017	2018	2019
$\overline{t}_{Y}^{HT,R}$	all	CZ	–0.03	–0.01	–0.01	–0.01
	all	SK	–0.01	0.01	0.00	0.02
$\overline{t}_{Y,(m,M)}^{HT,R}$	low	CZ	–3.36	–8.20	–16.0	–22.8
	medium		–2.04	–6.92	–14.8	–21.7
	high		–1.32	–6.23	–14.2	–21.2
	low	SK	–10.0	–4.21	–10.3	–15.5
	medium		–8.82	–2.88	–9.10	–14.3
	high		–8.14	–2.16	–8.44	–13.7
$\overline{t}_{Y,0.35,(m,M)}^{HT,R}$	low	CZ	–3.38	–3.61	–4.01	–4.91
	medium		–2.06	–2.17	–2.32	–2.82
	high		–1.34	–1.37	–1.45	–1.75
	low	SK	–5.23	–4.07	–4.65	–4.71
	medium		–3.22	–2.38	–2.78	–2.72
	high		–2.08	–1.50	–1.74	–1.66
$\overline{t}_{Y,0.50,(m,M)}^{HT,R}$	low	CZ	–3.38	–3.60	–4.01	–4.90
	medium		–2.06	–2.16	–2.32	–2.81
	high		–1.34	–1.37	–1.45	–1.74
	low	SK	–5.24	–4.07	–4.65	–4.71
	medium		–3.22	–2.38	–2.79	–2.72
	high		–2.08	–1.50	–1.75	–1.65

Source: Own construction

Table 7 Median absolute percentage error (MdAPE) – strategy S1

Estimate	Strategy for *(m, M)*	Country	Year			
			2016	2017	2018	2019
$\overline{t}_{Y}^{HT,R}$	all	CZ	1.14	1.10	1.03	1.03
	all	SK	1.29	1.26	1.23	1.14
$\overline{t}_{Y,(m,M)}^{HT,R}$	low	CZ	3.37	8.17	16.03	22.83
	medium		2.40	6.90	14.84	21.73
	high		2.34	6.28	14.30	21.25
	low	SK	10.05	4.22	10.30	15.52
	medium		8.82	3.00	9.08	14.36
	high		8.15	2.68	8.49	13.74
$\overline{t}_{Y,0.35,(m,M)}^{HT,R}$	low	CZ	3.36	3.58	4.02	4.91
	medium		2.24	2.24	2.37	2.82
	high		1.98	1.95	1.89	1.99
	low	SK	5.25	4.08	4.68	4.69
	medium		3.27	2.47	2.82	2.72
	high		2.37	2.02	2.14	2.02
$\overline{t}_{Y,0.50,(m,M)}^{HT,R}$	low	CZ	3.38	3.60	4.01	4.92
	medium		2.24	2.24	2.36	2.82
	high		1.97	1.89	1.84	1.93
	low	SK	5.26	4.09	4.68	5.99
	medium		3.28	2.47	2.82	2.70
	high		2.35	2.00	2.11	1.97

Source: Own construction

Table 8 Mean percentage error (MPE) – strategy S2

Estimate	Strategy for *(m, M)*	Country	Year			
			2016	2017	2018	2019
$\overline{t}_{Y}^{HT,R}$	all	CZ	–0.03	0.00	–0.01	0.00
	all	SK	–0.01	0.01	0.002	0.015
$\overline{t}_{Y,(m,M)}^{HT,R}$	low	CZ	–3.37	–8.11	–15.7	–22.4
	medium		–2.05	–6.83	–14.6	–21.3
	high		–1.32	–6.16	–14.0	–20.8
	low	SK	–10.0	–4.13	–10.1	–15.1
	medium		–8.81	–2.81	–8.89	–13.9
	high		–8.14	–2.08	–8.22	–13.3
$\overline{t}_{Y,0.35,(m,M)}^{HT,R}$	low	CZ	–3.38	–3.55	–3.94	–4.81
	medium		–2.06	–2.10	–2.24	–2.72
	high		–1.34	–1.31	–1.38	–1.65
	low	SK	–5.22	–3.90	–4.33	–4.35
	medium		–3.21	–2.23	–2.47	–2.36
	high		–2.06	–1.35	–1.42	–1.29
$\overline{t}_{Y,0.50,(m,M)}^{HT,R}$	low	CZ	–3.39	–3.55	–3.94	–4.80
	medium		–2.07	–2.09	–2.24	–2.72
	high		–1.35	–1.30	–1.38	–1.64
	low	SK	–5.22	–3.90	–4.33	–4.35
	medium		–3.21	–2.23	–2.48	–2.35
	high		–2.06	–1.35	–1.42	–1.29

Source: Own construction

Table 9 Median absolute percentage error (MdAPE) – strategy S2

Estimate	Strategy for *(m, M)*	Country	Year			
			2016	2017	2018	2019
$\overline{t}_{Y}^{HT,R}$	all	CZ	1.14	1.10	1.03	1.03
	all	SK	1.29	1.26	1.23	1.14
$\overline{t}_{Y,(m,M)}^{HT,R}$	low	CZ	3.37	8.17	16.03	22.83
	medium		2.4	6.90	14.84	21.73
	high		2.34	6.28	14.30	21.25
	low	SK	10.05	4.22	10.30	15.52
	medium		8.82	3.00	9.08	14.36
	high		8.15	2.68	8.49	13.74
$\overline{t}_{Y,0.35,(m,M)}^{HT,R}$	low	CZ	3.36	3.58	4.02	4.91
	medium		2.24	2.24	2.37	2.82
	high		1.98	1.95	1.89	1.99
	low	SK	5.25	4.08	4.68	4.69
	medium		3.27	2.47	2.82	2.72
	high		2.37	2.02	2.14	2.02
$\overline{t}_{Y,0.50,(m,M)}^{HT,R}$	low	CZ	3.38	3.60	4.01	4.92
	medium		2.24	2.24	2.36	2.82
	high		1.97	1.89	1.84	1.93
	low	SK	5.26	4.09	4.68	5.99
	medium		3.28	2.47	2.82	2.70
	high		2.35	2.00	2.11	1.97

Source: Own construction

CONCLUSION

The aim of the paper was to find rules of thumbs for parameters or RRT for surveys regularly conducted in time to study both mean and year-on-year growth rate. Study focused on the case of sensitive variable with broad range of values, therefore the proposed rules were evaluated on Czech and Slovak wage data coming from period of rapid growth in years 2016–2019. Since the aim of the paper is to find a general rule of thumb, the study excludes the year of Covid-19 pandemics. The economic shocks due to Covid-19 pandemics and corresponding measures on labour market would complicate finding this rule of thumb.

We apply methods using of dichotomous responses by Antoch et. al. (2022), because they are designed for sensitive variable with broad range of values and they overcome many limitations of standard methods using scramble variables (Vozár and Marek, 2022). We apply both original method and method using random numbers.

Two strategies combining three levels of the range (low, medium and high width) of interval (m, M) provided six possible rules of thumb. The S1 strategy is using the same interval in all years. The S2 strategy is based on changing lower bounds m using information on current minimum wage. The upper bound M was constant because there was no prior information for that. Updating by inflation rate had no practical effect because of low inflation in years 2016–2019. We choose the intervals by subjective expert idea what was low and high wage in this time. For methods using random numbers values of parameter α were set to recommended values 0.50 or 0.35 by Vozár (2024). As benchmark direct questioning under missing completely random and missing not at random responses model were also evaluated.

The conclusion of the simulation study is as follows. The heuristics behind the method using knowledge of random numbers works well. The reduction of variance and bias of population mean estimates is high, which supports conclusions from Vozár (2022). Values of parameter $\alpha = 0.50$ is a safe choice for any distribution of sensitive variable.

If the evolution of population mean in time is rapid, it is necessary to update the intervals (m, M) to avoid bias both in estimates of population mean and year-on-year growth rate. Interval bounds must be updated to avoid biases. The intervals (m, M) should be broad enough to cover as many units of population as possible. High variance and bias, if low non-response under missing not at random gives a convincing argument to use RRT instead of direct questioning. It is worth mentioning that the conclusion above was done by using both method on Czech and Slovak wage data in the period of growth. Therefore, further evaluation of more data sets and evolution patterns in time is recommended to refine the rule of thumb.

ACKNOWLEDGMENT

This work of was supported by the Institutional Support to Long-Term Conceptual Development of Research Organization, the Faculty of Informatics and Statistics of the University of Economics, Prague. The authors are grateful to two unknown reviewers for comments that considerably improved content of this paper.

References

ANTOCH, J., MOLA, F., VOZÁR, O. (2022). New Randomized Response Technique for Estimating the Population Total of a Quantitative Variable [online]. *Statistika: Statistics and Economy Journal*, 102(2): 205–227. <https://doi.org/10.54694/stat.2022.11>.

ARNAB, R. (1994). Nonnegative variance estimation in randomized response surveys [online]. *Communication in Statistics – Theory and Methods*, 23(6): 1743–1752. <https://doi.org/abs/10.1080/03610929408831351>.

BÍLKOVÁ, D. (2013). Modelování mzdových rozdělení posledních let v České republice s využitím L-momentů a predikce mzdových rozdělení podle odvětví. *E+M. Ekonomie a Management*, XVI(4): 42–54.

BÍLKOVÁ, D., MALÁ, I. (2012). Application of the L-Moment Method when Modelling the Income Distribution in the Czech Republic [online]. *Austrian Journal of Statistics*, 41(2): 125–132. <http://www.stat.tugraz.at/AJS/ausg122/122Bilkova1.pdf>.

CHAUDHURI, A., CHRISTOFIDES, T., RAO, C. (2016). *Handbook of Statistics 34. Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques*. Amsterdam: Elsevier. ISBN 978-0444-63570-9

DIANA, G., PERRI, P. F. (2013). Scrambled response models based on auxiliary variables. In: TORELLI, N., PESARI, F., BAR-HEN, A. (eds.) *Advances in Theoretical and Applied Statistics,* Berlin: Springer-Verlag, 281–291.

EICHHORN, B., HAYRE, L. S. (1983). Scramble randomized response methods for obtaining sensitive data. *Journal of Statistical Planning and Inference,* 7: 307–316.

ERIKSSON, S. (1973). A new model for randomized response [online]. *International Statistical Review*, 41: 101–113. <https://doi.org/10.2307/1402791>.

HYNDMAN, R. J., KOEHLE, A. B. (2006). Another look at measures of forecasts accuracy. *International Journal of Forecasting,* 22(4): 679–688.

JAKCSON, C. (2016). Flexsurv: a platform for parametric modelling in R. *Journal of Statistical Software*, 70: 1–33.

MALÁ, I. (2013). Použití konečných směsí logaritmicko-normálních rozdělení pro modelování příjmů českých domácností. *Politická ekonomie*, 61(3): 356–372.

MALÁ, I. (2015). Vícerozměrný pravděpodobnostní model rozdělení příjmů českých domácností. *Politická ekonomie*, 63(7): 895–908.

MAREK, L. (2010). Analýza vývoje mezd v ČR v letech 1995–2008. *Politická ekonomie*, 58(2):186–206.

MAREK, L., VRABEC, M. (2016). Using mixture density functions for modelling of wage distributions. *Central European Journal of Operations Research*, 24(4): 389–405.

R CORE TEAM (2024). *R: a language and environment for statistical computing.* Austria, Vienna: R Foundation for Statistical Computing.

TILLÉ, Y. (2006). *Sampling Algorithms*. New York: Springer. ISBN 978-0387-30814-2

VOZÁR, O. (2023). Unbiased variance estimator of the randomised response techniques for population mean [online]. *Statistika: Statistics and Economy Journal*, 103(1): 113–120. <https://doi.org/10.54694/stat.2022.38>.

VOZÁR, O., MAREK, L. (2023). Multicriteria Evaluation of Randomized Response Techniques for Population Mean [online]. *Statistika: Statistics and Economy Journal*, 103(4): 492–503. <https://doi.org/10.54694/stat.2023.32>.

VOZÁR, O. (2024). *Randomized Response Techniques for Population Mean*. Dissertation, Prague University of Economics and Business.

WARNER, S. (1965). Randomized response: a survey technique for eliminating evasive answer bias [online]. *Journal of American Statistical Association*, 60: 63–69. <https://doi.org/10.2307/2283137>.