HOW TO MAKE ESTIMATES WITH COMPENSATION FOR NONRESPONSE IN STATISTICAL ANALYSIS OF CENSUS DATA

Milan Terek
School of Management in Bratislava, Slovakia

Eva Muchova
University of Economics in Bratislava, Slovakia

Peter Lesko
University of Economics in Bratislava, Slovakia

ABSTRACT

The paper deals with the issue of solving nonresponse problems in a realized census. The purpose is to discuss the statistical methods we explain. To test the new approach, the data from the survey at one University are used. The suggested approach offers more accurate estimates because of compensation for nonresponse and the possibility to formulate broader conclusions based on the census data. The approach is advised in all surveys in which the costs of realization in the survey by census are practically the same as for sample survey, and the list of all units of the population is available.

Keywords: nonresponse in census, response propensity, poststratification using weights, correlation ratio.

DOI: http://dx.doi.org/10.15549/jeecar.v8i2.619

INTRODUCTION

At the academic year 2019/2020, a questionnaire survey was planned at the University of Economics in Bratislava (Slovak Republic). The aim of the project was to develop a multimedia and interactive framework for the teaching of the subjects Economic Theory 1 (ET1), and Economic Theory 2 (ET2), which are part of study programs at the first level of study. The sample survey and census were available. In both cases, it was necessary to cope with a higher nonresponse rate, which is common in current surveys.

When the nonresponse in the survey is only moderate, the accuracy of estimates should not be significantly influenced. A high level of nonresponse can significantly impair the quality and reporting capacity of the survey results (Cochran, 1977; Levy and Lemeshow, 2008; Lohr, 2010; Chaudhuri, 2014; Tille, 2020). In general, two types of nonresponse may be considered: unit nonresponse, which lacks the values of all

variables in the questionnaire and item nonresponse referring that the value of at least one but not all variables in the questionnaire are missing (Särndal and Lundström, 2005). Both types of nonresponse reduce the accuracy of estimates but are generally difficult to avoid.

Suppose the same response rate in sample survey and census. The number of responses in census is then higher. If the simple connection to all units of the population is available (Semin and Kislitskiy, 2020; Wonchan Ra, 2020) and if the costs of realization of the survey by census are practically the same as for sample survey, the sample survey does not make sense because of the unnecessarily loss of potential respondents. It is appropriate to send the questionnaires to all addresses in the database. This is how it was also done in the survey at the University of Economics.

There are many studies dedicated to nonresponse problem solving in sample surveys (Levy and Lemeshow, 2008; Lohr, 2010; Chaudhuri, 2014; Tille, 2020). These are focused on the methods of minimizing the negative impact of nonresponse to the accuracy of estimates. The question, how to make estimates with compensation for nonresponse in statistical analyses of census data, is studied and discussed in this paper. The purpose is to modify one method of estimation to compensate for nonresponse known in sample surveys, for use in censuses. The suggested approach offers more accurate estimates and the possibility to make broader conclusions, based on census data. The approach is advised in all surveys in which the costs of realizing the survey by census is not too different from a sample survey and the list of all population units is available.

Firstly, the sampling weights, response propensity, poststratification using weights and estimation using final weights in sample surveys will be described. Then the modification of poststratification using weights based on census data will be studied. In the framework of this, the question of which poststratification variables should be used in poststratification is also studied. The using of suggested approach will be tested on the data from a census realized at the University of Economics in Bratislava.

REVIEW OF LITERATURE

The effect of nonresponse on accuracy of estimates based on probability sampling is studied for example in Levy and Lemeshow (2008). The authors conclude that the bias given by non-response is independent of sample size n and cannot be reduced by its increasing. But it can be reduced by decreasing the proportion of units who would not respond, if selected. This indicates the great importance of preventative measures to reduce the proportion of units that would not respond. If the nonresponse rate is not negligible, inference based only upon the respondents may be seriously flawed.

In a probability sample, each unit in the population has a known probability of appearing in our selected sample. The probabilities π_i ,

$\pi_i = P(\text{unit } i \text{ in sample})$

are called the inclusion probabilities and they are known before the survey commences for any sampling design (Lohr, 2010). Sampling weights (base weights) w_{Bi} for any sampling design are defined as follows

$$w_{Bi} = \frac{1}{\pi_i}$$

The sampling weight of unit i can be interpreted as the number of population units represented by unit i.

The target population is the population to be studied in the survey and for which the basic inferences from the survey will be made. A frame population is a list compiled for the purpose of making a sample, which identifies the units of the population so that they can be taken into account when examining them (STN ISO 3534-1, 2008). Ideally, the target and frame population are identical. In practice, they are usually more or less different. The units that are in the target population but not in the frame population create the non-coverage.

Suppose a sample of n units is selected with known inclusion probabilities π_i , $i = 1, ..., n$. Frequently, the final survey weight, say w_i for observation i, is the product of three weight components. The first components w_{Bi} are derived from the survey design. The other weight components are regarded mainly as adjustments to the base weight. There are two types of adjustments: one to compensate for nonresponse, w_{NRi} , and one to compensate for

non-coverage, w_{NCi} . Thus, the final weight for the ith observation is the product of three components, i.e.

$w_i = w_{Bi} \cdot w_{NRi} \cdot w_{NCi}$

where w_{NRi} is the nonresponse adjustment factor and w_{NCi} is the poststratification adjustment factor using mainly for non-coverage compensation. For weights we will use, the following properties hold:

1. $\sum_{i=1}^{n_R} w_{Bi} = N_R$
2. $\sum_{i=1}^{n_R} w_{Bi} \cdot w_{NRi} = N_F$
3. $\sum_{i=1}^{n_R} w_i = \sum_{i=1}^{n_R} w_{Bi} \cdot w_{NRi} \cdot w_{NCi} = N$

where n_R is the sample size from the respondent population of the size N_R (Respondent population is the subset of frame population consisting of the units who would respond, if selected. It is only a theoretical concept), N_F is the frame population size and N is the target population size.

Let Z_i be the indicator variable for presence in the selected sample, with $P(Z_i = 1) = \pi_i$. Define the random variable R_i:

$$R_i = \begin{cases} 1 \text{ if unit } i \text{ responds} \\ 0 \text{ if unit } i \text{ does not respond.} \end{cases}$$

After sampling, the realizations of the response indicator variable R_i are known for the units selected in the sample. Let y_i equal a response of interest. A value for y_i is recorded if r_i , the realization of R_i, is 1. The probability that a unit selected for the sample will respond φ_i,

$\varphi_i = P(R_i = 1)$,

is unknown but assumed positive. The probability φ_i is called the response propensity for the ith unit. If R_i is independent of Z_i, then the probability that unit i will be measured is

$P(\text{unit } i \text{ is selected in sample and responds}) = \pi_i \varphi_i$.

The response propensity φ_i, is estimated for each unit in the sample, using auxiliary information that is known for all units in the selected sample. The final weight for a respondent is then $1/(\pi_i \hat{\varphi}_i)$, where $\hat{\varphi}_i$ is the estimated response propensity. Let $\boldsymbol{x}_i$ equal a vector of information known about unit i in the sample. If, φ_i depends on $\boldsymbol{x}_i$ but not on y_i, the data are missing at random (MAR data). For more details about MAR data, Missing Completely at Random (MCAR data) and Not Missing at Random (NMAR data), see Lohr (2010). There are several ways to estimate the response propensity φ_i for each unit in the sample. The simplest approach is the application of weighting methods. One of them is poststratification using weights. Alternatively, logistic regression (see Montgomery et al., 2012, Larose, 2006) can be used to estimate the response propensity (see Levy and Lemeshow, 2008). An important disadvantage of the application of this method is that the nonresponse adjustment factors may be quite unstable, which can lead to widely varying and extreme weights. When extreme weights are present, their trimming is advised (Potter, 1988, 1990).

METHODS

Poststratification Using Weights

Stratification is the distribution of a population into sub-sets (strata), which are mutually exclusive and completely covering the population. The strata are considered more homogeneous with respect to the studied variable than the whole population (STN ISO 3534-1, 2008). When the stratification is performed after simple random sampling, it is a poststratification. Both stratification and poststratification are based on auxiliary information. The variables that serve as a criterion for stratification (poststratification) are most often called stratification (poststratification) variables. They should strongly correlate with the studied variables. Poststratification requires auxiliary information - distribution of the population according to poststratification variables (Lohr, 2010).

The basic idea to use poststratification to try to compensate for nonresponse is dividing the sample into groups based on variables that are known for both respondents and nonrespondents and are thought to be related to response propensity. The response rate for the poststrata is taken as the response propensity for each sample member in the poststratum. It is assumed that responding and non-responding units in the same poststrata are similar. The weights of responding units from the same poststrata are increased, so that they represent

also the non-responding units. We modify the base weights so that the sample is calibrated to population counts in the poststrata. Poststratification is a special case of calibration methods in survey sampling (Deville and Särndal,1992; Särndal, 2007).

Suppose a simple random sample is taken. After the sample is collected, units are grouped into H different poststrata. The population has N_h units in hth poststratum; of these, n_h were selected for the sample and n_{hR} responded. The response propensity for each unit in poststratum h (h = 1, 2, …, H) will be estimated by weighted response rate. For every respondent i in poststratum h, the response propensity is estimated by RR_{w_h},

$$RR_{w_h} = \frac{\sum_{i=1}^{n_{hR}} w_{Bi}}{N_h} \quad (1)$$

and nonresponse adjustment factor is

$$w_{NRi} = \frac{1}{RR_{w_h}} \quad (2)$$

When w_{NRi} > 2, the poststratum contains more non-respondents than respondents. In such cases, the variance of the estimator increases; the weights may not be stable. The collapsing of neighboring poststrata is advised to obtain nonresponse adjustment factors of 2 or less (Lohr, 2010). The same is advised when the number of observations in a poststrata is less than 20. In Gelman and Carlin (2002) it is advised to collapse the poststrata with others that have similar means in key variables until they have a reasonable number of observations in each poststratum. Other authors studied the same problem in weighting class adjustment method (WCA). In Eltinge and Yansaneh (1997) the methods for choosing the number of weighting classes to use are discussed. In Little and Vartivarian (2003) and Vartivarian and Little (2003) is suggested that it may be inefficient to use weighted response rates as estimates of response propensities for WCA cells. Rather they recommend incorporating survey design variables (such as stratification variables) in the definition of WCA cells as well as variables that are related to response propensity and the survey outcomes. When poststrata are formed using more than one variable, but only the marginal population totals are known, the ranking adjustments method can be used to adjust for nonresponse and under-coverage (Brackstone and Rao, 1979). The algorithm may not converge if some of the cell estimates are zero and there is also a danger of "overadjustment".

Estimation Based on Final Weights

We consider y_i to be a measurement on observation unit i, and w_i to be the final weight of observation unit i. The particular chosen sample will be denoted by S, a subset consisting of n of the units in the population U. The general estimator of the population total $\hat{\tau}$ is (Lohr, 2010)

$$\hat{\tau} = \sum_{i \in S} w_i y_i \quad (3)$$

where all measurements are at the observation unit level.

The general estimator of the population mean $\hat{\mu}_K$ is

$$\hat{\mu}_K = \frac{1}{\sum_{i \in S} w_i} \sum_{i \in S} w_i y_i, \quad (4)$$

where $\sum_{i \in S} w_i$ estimates the number of observation units in the population.

Define y_i to be 1 if the unit has the characteristic and to be 0 if the unit does not have that characteristic. Then the proportion π is

$$\pi = \frac{\sum_{i=1}^{N} y_i}{N} \quad (5)$$

and π is estimated by $\hat{\pi} = \hat{\mu}_K$.

The poststratified estimator of the mean or total is approximately unbiased if within each poststratum h, (a) the response y_i is uncorrelated with the response propensity φ_i , (b) the response propensity φ_i is the same for every unit, or (c) the value of y_i is the same. The using of many poststrata is advised, to make the meeting of these conditions the most plausible (Lohr, 2010).

NONRESPONSE IN CENSUS

Until now we considered some sampling design containing the simple random sampling. In a census of the population of the size N, all units from the population are selected. We can understand it as random sampling without replacement of size N. The only difference is that the last unit is selected non-randomly. The

inclusion probability of a unit in such a sample (census) is equal to one, equally as its base sampling weight.

In census, weights can also be used to adjust for nonresponse. The random variable R_i is defined identically as in the sample survey. The probability that a unit i will respond φ_i,

$\varphi_i = P(R_i = 1)$,

is unknown but assumed positive. The probability φ_i is the response propensity for the ith unit also in the census. In a census, the probability that unit i is selected and responds is reduced to the probability that unit i responds because all units from the population are selected. So

P(unit i selected and responds) = φ_i

The probability of responding, φ_i is estimated by $\hat{\varphi}_i$ for each unit in the population, using auxiliary information that is known for all units in the population. The final weight for a respondent is then $1/\hat{\varphi}_i$. We will assume MAR data. Then poststratification using weights can be used.

The population has N_h units in poststratum h; of these, N_{hR} responded. The response propensity for each unit in poststratum h will be estimated by weighted response rate. Then for every respondent i in poststratum h, the response propensity is estimated by $\hat{\varphi}_i$,

$$\hat{\varphi}_i = RR_{w_h} = \frac{\sum_{i=1}^{N_{hR}} w_{Bi}}{N_h} = \frac{N_{hR}}{N_h} \qquad (6)$$

and the final weight w_i for unit i is

$$w_i = \frac{1}{RR_{w_h}} = \frac{N_h}{N_{hR}} \qquad (7)$$

In a census, the final weight is equal to the nonresponse adjustment factor (see also Terek, 2020).

If the decision to apply this method was taken, the question is, which poststratification variables should be used in poststratification? Usually, accessible auxiliary information offers more potential poststratification variables. It is known that the bias of estimators caused by nonresponse can be minimized by finding poststratification variables that are highly correlated with the response propensity. As a suitable tool for measuring the strength of the relationship between the response propensity and poststratification variables, we propose a correlation ratio $\eta_{(Z|X)}$. Correlation ratio is a measure of the relationship between the statistical dispersion within individual categories and the dispersion across the whole population or sample. It is defined as the ratio of two standard deviations representing these two types of variation.

Suppose each observation (value of quantitative variable) is z_{xi} where x indicates the category that observation is in and i is the index of the particular observation. Let n_x be the number of observations in category x and

$$\bar{z}_x = \frac{\sum_i z_{xi}}{n_x} \qquad (8)$$

and

$$\bar{z} = \frac{\sum_x n_x \bar{z}_x}{\sum_x n_x} \qquad (9)$$

where $\bar{z}_x$ is the mean of the category x and $\bar{z}$ is the mean of the whole population. The correlation ratio $\eta_{(Z|X)}$ is defined as to satisfy

$$\eta^2_{z|x} = \frac{\sum_x n_x (\bar{z}_x - \bar{z})^2}{\sum_{x,i} \left(z_{x,i} - \bar{z}\right)^2} \qquad (10)$$

The correlation ratio takes values between 0 and 1. The limit $\eta_{z|x} = 0$ represents the case of no dispersion among the means of the different categories, while $\eta_{z|x}$ = 1 refers to no dispersion within the respective categories. In the context of nonresponse analysis, the variable z will be quantitative discrete variable taking two values 1 and 0 – the number of responses. If the respondent i from category x answered, z_{xi} = 1, if not, z_{xi} = 0.

Nonresponse in census data from the survey realized at University of Economics

The mentioned census was realized at the end of the academic year 2019/2020, within the project "Learn Economics", concerning the teaching of the subjects Economic Theory 1 (ET1) and Economic Theory 2 (ET2). The survey was conducted through Google forms software. The contingency tables were created and analyzed with Microsoft Excel. A total of 1,351 students of the University of Economics, completed at least one of the subjects in the given academic year. We have the e-mail addresses of all students who

passed the exam ET1 and (or) ET2 during the academic year 2019/2020. A questionnaire was sent to each of them. It consisted of 36 questions focused on the content, methodology, form of lectures and seminars as well as on the evaluation of online teaching, which was applied on the teaching of ET2 in the summer semester due to the epidemiological situation. The questionnaire contained 3 closed and 33 open questions. From them only closed questions were interesting for quantitative analysis, and this included in the analysis within this study. The completed questionnaire was returned by 429 students, so the total response rate was 429/1351 = 31.75%. In addition to the information from the questionnaires, we also obtained some auxiliary information about the population of 1,351 students, from study department of University of Economics. Specifically, the frequency distribution by faculty and year of study and frequency distribution by faculty and gender are in Tables 1 and 2 (in the parentheses are the corresponding numbers of responding students). It is interesting in Tables 1 and 2, that differences in response propensity are maximal between man and women (range R = 0.378 – 0.306 = 0.137), among faculties are less (range R = 0.372 – 0.260 = 0.112) and between first and second year of study is minimal difference (R = 0.315 – 0.306 = 0.009. We did not take in consideration third year of study because of too small number of respondents).

Table 1. Distribution of students who completed at least one of the subjects in the given academic year, by faculty and year of study

Faculty Year of study	Faculty of National Economy	Faculty of Business Informatics	Faculty of Commerce	Faculty of Business Management	Faculty of International Relations	Faculty of Applied Languages	Total
1.	275 (91)	236 (78)	248 (63)	4 (0)	127 (47)	33 (10)	923 (289)
2.	51 (11)	36 (7)	60 (15)	223 (79)	42 (16)	0 (0)	412 (128)
3.	4 (2)	2 (2)	3 (3)	3 (3)	3 (1)	1 (1)	16 (12)
Total	330 (104)	274 (87)	311 (81)	230 (82)	172 (64)	34 (11)	1351 (429)

Source: Developed by authors

Table 2. Distribution of students who completed at least one of the subjects in the given academic year, by faculty and gender

Faculty Gender	Faculty of National Economy	Faculty of Business Informatics	Faculty of Commerce	Faculty of Business Management	Faculty of International Relations	Faculty of Applied Languages	Total
Males	137 (29)	123 (29)	135 (30)	93 (20)	66 (19)	5 (3)	559 (130)
Females	193 (75)	151 (58)	176 (51)	137 (62)	106 (45)	29 (8)	792 (299)
Total	330 (104)	274 (87)	311 (81)	230 (82)	172 (64)	34 (11)	1351 (429)

Source: Developed by authors

Now, we will calculate the correlation ratio measuring strength of the relationship between the response propensity and variables Faculty and Gender. For responding student i from category x, z_{xi} = 1, for non-responding student, z_{xi} = 0. Firstly, the averages $\bar{z}_x$ for each category x are calculated. For example, for category Males – Faculty of National Economy, the average number of responses per one addressed student (proportion of responding students) is 29/137 = 0.212. For category Males – Faculty of Business Informatics, the average number of responses per one addressed student is 29/123 = 0.236, and so on. The grand average $\bar{z}$ is 429/1351 = 0.318. Now, we can substitute for averages $\bar{z}_x$ and grand average $\bar{z}$, to relation (10), calculate $\eta^2_{z|x}$ and from it, $\eta_{(z|x)}$ = 0.221. Similarly, the values of $\eta_{(z|x)}$ for other variables, we have the data about, can be calculated. The all results are in Table 3.

Table 3 Values of correlation ratio

Variables	$\eta_{(z\|x)}$
Faculty	0,096
Year of study	0,119
Gender	0,178
Faculty – Gender	0,221
Faculty – Year of study	0,190

Source: Developed by authors

It can be seen in Table 3, that the Faculty - Gender have maximum correlation ratio and so, they will serve as poststratification variables. So, all combinations of categories of Faculty and Gender define poststrata. Originally we have 2 x 6 = 12 poststrata.

We will analyse in details the closed question number 29 to prove our original proposition as discussed in the preceeding sections of this study. The question posed to students were "*How do you evaluate the using of Moodle in teaching ET1/ET2*"? The possible answers are – "positive", "rather positive", "rather negative", "negative". The 423 students answered this question. According to the requirement to have in each poststrata at least 20 responding units, we collapsed two last columns in Table 2. After collapsing we have 2 x 5 = 10 poststrata. The resulting structure of poststrata with the number addressed and responding students (in parantheses) is in Table 4. Comparing Tables 2 and 4 it is clear, that 3 men and 3 women who returned the completed questionnaire did not answer this question.

Table 4. Structure of poststrata with the number of responding students in each of them

Faculty Gender	Faculty of National Economy	Faculty of Business Informatics	Faculty of Commerce	Faculty of Business Management	Faculties of International Relations and Applied Languages	Total
Males	137 (28)	123 (28)	135 (29)	93 (20)	71 (22)	559 (127)
Females	193 (75)	151 (58)	176 (50)	137 (60)	135 (53)	792 (296)
Total	330 (103)	274 (86)	311 (79)	230 (80)	206 (75)	1351 (423)

Source: Developed by authors

In Table 5 are the final weights calculated from the data in Table 4, according to relation (7). For example, w_{11} = 137/28 = 4.893; w_{12} = 123/28 = 4.393, and so on. With aid of final weights, the proportions of particular answers can be estimated by relation (4). Firstly, we will estimate the proportion of answering "positive". For it we need the contingency table with the number of answers "positive" in each poststratum (in Table 6).

Table 5. Final weights

Faculty Gender	Faculty of National Economy	Faculty of Business Informatics	Faculty of Commerce	Faculty of Business Management	Faculties of International Relations and Applied Languages
Males	4.893	4.393	4.655	4.650	3.227
Females	2.573	2.603	3.520	2.283	2.547

Source: Developed by authors

Table 6. Number of answers "positive" in poststrata

Faculty Gender	Faculty of National Economy	Faculty of Business Informatics	Faculty of Commerce	Faculty of Business Management	Faculties of International Relations and Applied Languages
Males	15	20	9	7	12
Females	41	32	20	37	27

Source: Developed by authors

Then we can calculate, based on data from tables 5 and 6, the value of nominator of relation (4).

$\sum_{i\epsilon S} w_i y_i$ = 4.893 x 15 + 4.393 x 20 +...+ 2.547 x 27 = 686.853

and

$\sum_{i\in S} w_i$ = 1351.

Then

$\hat{\pi}_{positive} = \frac{686.853}{1351}$ = 50.84%.

Similarly, we can calculate

$\hat{\pi}_{rather\ positive}$ = 43.51%; $\hat{\pi}_{rather\ negative}$ = 3.90%; $\hat{\pi}_{negative}$ = 1.75%.

How can we interpret the obtained results? The estimates compensating for nonresponse were obtained. The results can be interpreted as follow: We estimate that 50.84% of 1,351 students who completed at least one of the subjects ET1/ET2 in the given academic year, evaluate the using of Moodle in teaching as positive, 43.51% as rather positive, 3.90% as rather negative and 1.75% as negative. The normally used procedure is such that we simply compute proportions of the answers of responding students. We know that 423 students answered this question. From them, 220 answered" positive", 181 "rather positive", 16 "rather negative" and, 6 "negative". So, proportions are $p_{positive}$ = 220/432 = 52%, $p_{rather\ positive}$ = 181/423 = 42.79%, $p_{rather\ negative}$ = 16/423 = 3.78%, and $p_{negative}$ = 6/423 = 1.42%. These results can be interpreted as follows:

From 1,351 addressed students, 423 answered this question. From them 52% answered "positive", 42.79%, "rather positive", 3.78% "rather negative" and, 1.42% "negative". We did not say anything about the population of 1,351 students, we characterized only self-selection of 423 responding students.

The closed question number 13 is "*Did you use tutoring during your studies of ET1 and (or) ET2*"? The possible answers are – „yes" and „no". The question was answered by 427 students. The weights were calculated and then by the same way as before, the estimate of "Yes responses" was computed: $\hat{\pi}_{yes}$ = 10.99%. We estimate that 10.99% of 1,351 students which completed at least one of the subjects ET1/ET2 in the given academic year, used tutoring during their studies. For comparison, from 427 students who responded, 48 made choice „yes", so p_{yes} = 11.24%. From 1,351 addressed students, 427 answered this question. From them 11.24% answered "yes". This result also illustrates the qualitative difference of information obtained based on $\hat{\pi}_{yes}$ and on p_{yes}.

The closed question number 24 is "*Did you use during semester online platform Moodle*?" The possible answers are – „yes" and „no".The question answered 425 students. The estimate of "Yes responses" is $\hat{\pi}_{yes}$ = 96.41 %. We estimate that 96.41% of 1,351 students which completed at least one of the subjects ET1/ET2 in the given academic year, used during semester online platform Moodle. For comparison, from 425 students who responded, 412 made choice „yes", so p_{yes} = 96.94%. From 1,351 addressed students, 425 answered this question. From them 96.94% answered „yes".

The difference between two approaches is evident. There are differences in results as well as in interpretation power. If we take, for example, the answers to question 29, the absolute values of differences between corresponding values of $\hat{\pi}$ and p, are accordingly 1.16%, 0.72%, 0.12% and 0.33%. Because of compensation for nonresponse, the suggested approach offers more accurate results. In addition, it estimates the quantities of the population, while the normally used procedure only describes self-selection of responding units.

CONCLUSIONS

There are a lot of cases, when analysts can simply contact all population units, the difference between costs of census and sample survey is minimal or does not exist, and some useful auxiliary information about population units is available. This is also the case of presented census. Then it does not make sense to realize a sample survey. The census is more suitable because no potential respondent is left out. The purpose of the paper was to propose the modification of one method of estimation with compensation for nonresponse known in sample surveys, for its use in censuses. The proposed modification enables to compensate for nonresponse in estimation based on census data and to make broader conclusions about the studied population.

It is known that the bias of estimators caused by nonresponse can be minimized by finding poststratification variables that are highly correlated with the response propensity. We propose as a suitable tool for measuring the strength of the relationship between the response propensity and poststratification variables, a correlation ratio. In the presented census for the variables Faculty – Gender is the correlation ratio maximal (in Table 3), so these variables creating 12 poststrata are the poststratification variables. If the response rate for each poststratum is not at least 50% or the number of observations in each poststratum is not at least 20 observations, the weights can be unstable. In the presented census we succeed to meet only the second requirement by collapsing of two faculties, so a less stability of weights is possible. The comparison of the proposed and normally used procedure shows the differences in results. The proposed approach offers the estimates of population quantities taking into account also nonresponse, while the normally used procedure only simply describes self-selection of responding units.

The identical survey at University of Economics is planned to be yearly repeated also in the future, the results will be compared, and the development mapped. In the future studies, the analysis of association between two categorical variables could be of interest. When simple random sample is disponible, the structure of association can be revealed by adjusted

standardized residuals. How can we reveal the structure of association based on census data? The possibilities of using the odds ratio for measuring the strength of association based on census data could be also interesting.

ACKNOWLEDGEMENTS

This paper was supported by grant from Grant Agency KEGA in the framework of the project 035EU-4/2020 „Learn Economics: Application of E-learning as a New Form of Teaching of Economics".

REFERENCES

Brackstone, G. J., and Rao, J. N. K. (1979). An investigation of raking ratio estimators. *Sankhya, Series C*, *41*, pp. 97–114

Chaudhuri, A. (2014). *Modern Survey Sampling*. Broken Sound Parkway NW: CRC Press

Cochran, W. G. (1977). *Sampling Techniques*. *Third edition*. New York: J. Wiley and Sons

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, pp. 376–382

Eltinge, J. L., and Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, pp. 33–40

Gelman, A., and Carlin, J. B. (2002). Poststratification and weighting adjustments, R. M. Groves, D. Dillman, J. Eltinge, and R. Little (Eds.). *Survey nonresponse*, New York: Wiley and Sons, pp. 289–302

Larose, D.T. (2006). *Data Mining Methods and Models*, Hoboken: Wiley and Sons

Levy, P. S., and Lemeshow, S. (2008). *Sampling of Populations. Methods and Applications, Fourth Edition*, Hoboken: Wiley and Sons

Little, R. J., and Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 22, pp. 1589-1599

Lohr, S. L. (2010). *Sampling: Design and Analysis, Second Edition*, Boston: Brooks/Cole

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2012). *Introduction to Linear Regression Analysis, Fifth Edition*, Hoboken: Wiley and Sons

Potter, F. (1988). Survey of procedures to control extreme sampling weights, *ASA Proceedings of the Survey Research Methods Section,* pp. 453-458

Potter, F. (1990). A Study of procedures to identify and trim extreme sampling weights, *Proceedings of the Survey Research Methods Section,* pp. 225-230

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Hoboken: Wiley and Sons

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, *33*, pp. 99–119.

Semin,A, and Kislitskiy, M.(2020). The Econometric Model for Assessing the Economic Category of a Russian Farmer Entrepreneur in Terms of the "Innovator vs. Conservative" System. *Journal of Eastern European and Central Asian Research*, Vol 7 No 3

STN ISO 3534-1 (2008). Statistika. Slovnik a znacky. Cast 1: Vseobecne statisticke terminy a terminy pouzivane v teorii pravdepodobnosti. Bratislava: Slovensky ustav technickej normalizacie 2008

Terek, M. (2020). Moznosti riesenia problemu neodpovedania v analyzach dat pri vycerpavajucom skumani prostrednictvom dotaznikovych zistovani. *Slovenska statistika a demografia* 4/2020

Tille, Y. (2020). *Sampling and Estimation from Finite Populations.* Hoboken: Wiley and Sons

Vartivarian, and Little, R. (2003). On the formation of weighting adjustment cells for unit nonresponse. In *Proceedings of the Survey Research Methods Section*, American Statistical Association

Wonchan Ra (2020). Determinants of the Choice of Combined Modes for Foreign Market Entry: The Case of Korean Firms Entering into Uzbekistan, *Journal of Eastern European and Central Asian Research*, Vol 7 No 1

ABOUT THE AUTHORS

Milan Terek, email: milan.terek1@gmail.com

Milan Terek, PhD. works from 2018 as full professor at School of Management in Bratislava (course leader on: Introduction to Statistics, Statistics, Mathematics for Managers II, Quantitative Methods for Managers, Quantitative methods in Business Management Research). Before he worked at University of Economics in Bratislava (course leader on: Statistics, Statistical Quality Control, Decision Analysis, Data Mining, Survey Sampling, Linear Programming, Nonlinear Programming, Operations Research, System Modelling). His research activities are oriented on the applications of statistical methods in economics and management.

Eva Muchova, PhD. is a full professor at the Department of Economics at the University of Economics in Bratislava, Slovakia. She is a course leader on Macroeconomics and International Economics. Her research is focused on the European integration, open economy macroeconomics, international relations and methodology of teaching economics.

Peter Lesko, PhD. is an assistant professor and researcher at the Department of Economic Theory, Faculty of National Economy, University of Economics in Bratislava (course leader on: Economic Theory 1, Economic Theory 2, Macroeconomics 2). His research activities are oriented on the issues of constrained economic growth in the open economy (Thirlwall` s law), the convergence sustainability of Central and Eastern European countries and the concept of sustainable economic development.