The impact of changes in sample allocation on distributions of examined sample variables

Janusz Niezgoda[1]

Abstract

The knowledge of the population may be taken from exhaustive or partial research. It is assumed that exhaustive research provides full information about the population under analysis. However, because of the constant increase of research costs, the short validity of data and the continuous demand for information, it is necessary to use the representative sampling method, allowing researchers to acquire knowledge quickly. Being based on the examination of a part of the population (sample), it brings forth such problems as the method of sampling, missing answers or an increase of the sample.

The aim of this study is to analyse the results of modifying the stratified sampling scheme, where the size of the sample is increased in selected strata.

The research was done with the use of computer simulation and the sampling scheme in stratified sampling.

The results show that an increase of the size of the population in selected strata results in a change of the distribution of the examined variable.

Keywords: stratified sampling scheme, random variable distribution, non-profit organisations

JEL Classification: C46, C83, L31

1. Introduction

From the viewpoint of science, it is best to examine the entire population. In the case of large populations, the basic shortcomings of such research (exhaustive research) are its long duration and high costs of performance. The protection of secrecy and the ordinary reluctance to provide necessary information also pose serious problems. In such a situation, the only option is to use the representative sampling method in research. The primary tools used in it are described thoroughly, among others, in works by (Steczkowski and Stefanów, 2009; Cochran, 1997; Zasępa, 1962).

Examining a statistical community by means of the representative sampling method certainly has such advantages as the shorter time of examination and lower costs. For these reasons, the representative sampling method is applied in social and economic sciences, e.g. in the case of research on households or public opinion research (Groves, 2006; Benade, 2019).

[1] Cracow University of Economics, Department of Statistics, januszni@uek.krakow.pl.

In computer science it is used, among others, for assessing the reliability of software (Podgurski, 1999) or analysing large data collections (Zhao, 2018). The major shortcoming of the representative sampling method is the presence of errors resulting from incomplete information about the population under analysis (Groves, 2006). Consequently, determined values of estimators are burdened with errors. For the purpose of their decrease, some researchers make attempts to increase the size of the sample in selected strata above the size resulting from the adopted stratified sampling scheme. However, such an action has an impact on the resulting distribution of sample variables.

The aim of this paper is to analyse the results of modification of the stratified sampling scheme, where the size of the sample is increased in selected strata.

2. Stratified sampling scheme

In the case of homogeneous communities, simple sampling schemes are usually used. The use of the stratified sampling scheme is recommended in the following situations:

a. when the population is diversified at least with regard to one variable;
b. when it is necessary to acquire information for some parts of the population (subpopulations), each part should be treated as the whole population;
c. administrative comfort may prescribe the use of stratification, for example, when an agency conducting a survey has local offices each of which can conduct a survey for a part of the population;
d. the diversification of the population causing various sampling problems in different parts of the population (Cochran, 1977, pp. 89-90).

If the population is not divided in a natural way, its stratification must be performed in a manner guaranteeing the fulfilment of two conditions: each element of the population is assigned to one stratum, and the sum of all elements constitutes the population. In stratification procedures there is also a requirement that particular strata be internally homogeneous and maximally diversified between one another Cochran (1977), Zhao (2019). Because of the division of the population, it becomes necessary to allocate the sample in strata (Steczkowski, 2009, pp. 74-80; Zhao, 2019, p. 419; Benade, 2019). There are four methods of sample allocation in strata:

- Uniform allocation, where equipotent samples are taken from each strata:

$$n = \sum_{h=1}^{l} n_h \tag{1}$$

where:
n – sample size n_h – sample size in h – this stratum l – number of strata

- Proportional allocation is based on sampling from each sample stratum in a manner ensuring the fulfilment of the condition:

$$\frac{N_h}{N} = \frac{n_h}{n} \tag{2}$$

where:
N – population size
N_h – stratum size

In the case of uniform and proportional allocation, it is first necessary to determine the minimum size of the sample:

$$n = \frac{u_a^2 s^2 N}{u_a^2 s^2 + N d^2} \tag{3}$$

where:
u□ – quantile of the normal distribution read for the confidence level equal to 1 – □
d – requested precision of evaluation
s^2 – variation of population from preliminary studies

- Apart from sizes n_h, Neyman allocation takes intra-stratum variances into consideration:

$$n_h = \frac{N_h s_h}{\sum_{h=1}^{l} N_h s_h} \tag{4}$$

where:
s_h – standard deviation in h – this stratum from preliminary studies

- Optimum allocation also takes into account the diversification of research costs in each stratum c_h:

$$n_h = \frac{\frac{N_h s_h}{\sqrt{c_h}}}{\sum_{h=1}^{l} \frac{N_h s_h}{\sqrt{c_h}}}. \tag{5}$$

Publications concerning the representative sampling method focus mainly on the properties of estimators.

3. Characteristics of the subject matter of research

The starting point for considerations is the sampling scheme in research on the Polish non-profit organisation sector described in [GUS (Central Statistical Office of Poland) 2016, pp. 22-24]. The current sampling scheme assumes the purposive-random sampling of research units. In the first place, the purposive sampling of units meeting specific conditions is carried out. Inter alia, all of the following units are selected:

- units whose business activity has been registered;
- units having the status of a public benefit organisation;
- originators of the European Social Fund;
- units employing more than five persons;
- water volunteer rescue service;
- mountain volunteer rescue services;
- Tatra Mountains Volunteer Rescue Service;
- units that have concluded a contract with the National Health Fund.

For other units included in the sampling frame, the sample taken has a representative nature within each province and type of organisation. The strata were provinces with separate cities having more than 500,000 inhabitants. The sampling for strata was not proportional.

The size of the sample was determined on the assumption that the relative standard error of estimated parameters should not exceed 5%. Due to the assumed 15% participation of inactive units, the number of units randomly sampled in each stratum was 20% higher.

4. Structure of the population under analysis

In this study only the sampling scheme for the population of foundations was analysed. For all provinces, the size distributions were built by grouping counties according to the number of foundations operating in them. In Table 1, for example, foundation number distributions were presented for the Mazovia and Opole provinces. This distributions based on the data of the Statistical Office in Kraków. Because all the provinces contained counties with largely outlying sizes, they were replaced with an artificial category 105-115 and a size selected so as to make the total number of foundations closest to the actual one.

5. Simulation experience and results

For the purpose of examining the impact of increasing the size of samples taken from individual strata, the provinces were divided into two groups: "strong" provinces, where larger and more developed foundations prevail (e.g. with a larger number of employees and higher revenues), and "weak" provinces with a prevalent number of small foundations (e.g. with lower revenues). The group of strong provinces consisted of: Kujawy-Pomerania, Łódź, Małopolska, Mazovia, Pomerania, Silesia, Lower Silesia and Wielkopolska. The group of poor provinces consisted of: West Pomerania, Warmia-Mazury, Świętokrzyskie, Podlasie, Podkarpacie, Opole, Lubuskie and Lublin.

Table 1. Structure of foundations in counties of the Mazovia and Opole provinces

j	Number of foundations		Mazovia Province	Opole Province
	x_{dj}	x_{gj}	f_j	f_j
1	0	10	19	7
2	10	20	7	2
3	20	30	4	2
4	30	40	1	0
5	40	50	1	0
6	50	60	2	0
7	60	70	2	0
8	105	115	47	1

Strong provinces were described by random variables: $X_{1.m} \sim N(20, 1)$ reflecting such continuous variables as revenues.

Dichotomous variables $X_{2.m}$ with a two-point distribution ($x_{2.m} = 1$ with probability $p = 0.2$; $x_{2.m} = 0$ with probability $p = 0.8$) describing, for example, the question from Section I of the SOF-1 form with numbers 1, 11 and 12. Multi-criterion answers, e.g. Section II, question 1, Section VII, question 2 of the SOF-1 form were simulated by a four-point distribution.

Weak provinces were described by random variables: $X_{1.s} \sim N(10, 1)$, dichotomous ($x_{2.m} = 1$ with probability $p = 0.8$; $x_{2.m} = 0$ with probability $p = 0.2$) and four-point distribution.

Data on distributions stated above were generated for each county, thus building the population. This paper presents preliminary results based on a single sample from the population. Its distributions are presented in Fig. 1, 2 and 3 (the population variant).

First, minimum sample sizes were determined for the provinces, being divided proportionally into counties. The distribution from the standard sample is presented in Fig. 1, 2 and 3 (the standard sample variant). Later, a modified sample was taken, where the minimum sample size was increased in counties where up to five foundations operate. The distributions

are presented in Fig. 1, 2 and 3 (the modified sample variant). Table 2 contains results of the Kolmogorov-Smirnov test, which is used for the verification of the hypothesis about the consistency of distributions of the variable X_1 in the population and both samples.

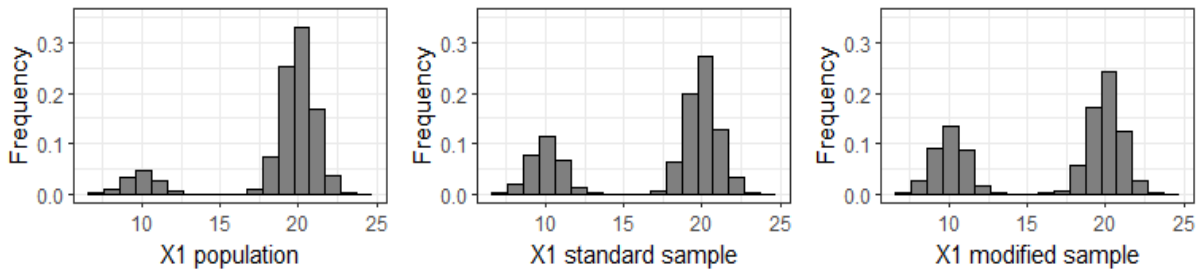

Figure 1. Distributions of continuous random variables

Table 2. Results of the Kolmogorov-Smirnov test for the continuous variable

Community	Standard sample	Modified sample
Population	D = 0.1737 p-value < 2.2e-16	D = 0.2442 p-value < 2.2e-16

Distributions of the dichotomous variable and the four-state variable are presented in Tables 3 and 5. It is easy to notice growing differences between the population and the standard sample on the one hand and the population and the modified sample on the other hand.

Table 3. Distributions of the dichotomous variable

	X_2	
	0	**1**
Population	73.54%	26.46%
Standard sample	63.46%	36.54%
Modified sample	58.90%	41.10%

Table 4. Results of the equivalence test of two structure indicators for the dichotomous variable

		$X_2 = 0$		$X_2 = 1$	
		Standard sample	Modified sample	Standard sample	Modified sample
Population	□2	208.74	388.81	208.24	388.09
	p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

This observation is confirmed by results of the equivalence test of two structure indicators presented in Table 4 and Tables 6 and 7.

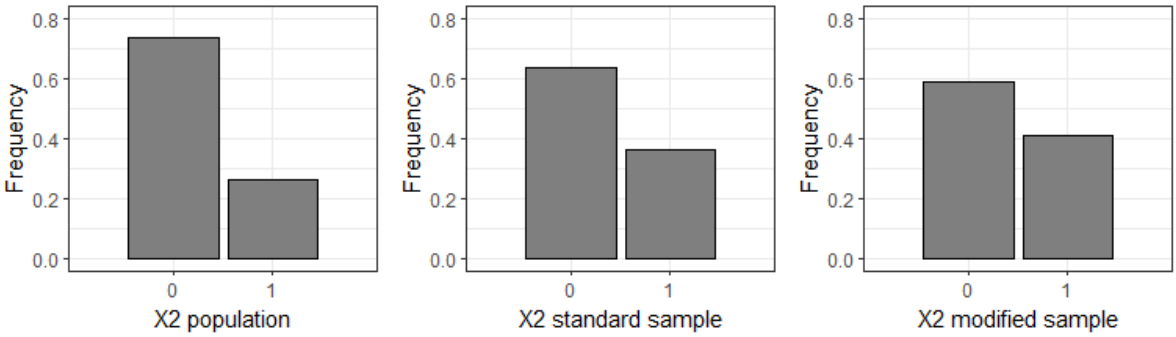

Figure 2. Distributions of dichotomous variables

Table 5. Distributions of the four-point variable

	X_3			
	0	**1**	**2**	**3**
Population	11.41%	22.83%	40.06%	25.71%
Standard sample	22.92%	23.95%	32.60%	20.53%
Modified sample	27.49%	24.11%	29.42%	18.98%

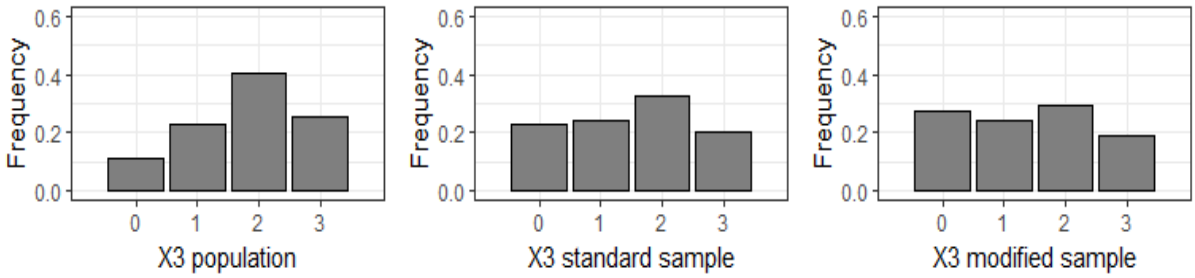

Figure 3. Distributions of the four-point variable

Table 6. Results of the equivalence test of two structure indicators for the four-point variable for $x_3 = 0$ and $x_3 = 1$

		$X_3 = 0$		$X_3 = 1$	
		Standard sample	Modified sample	Standard sample	Modified sample
Population	□2	471.5	796.56	2.905	3.3842
	p-value	< 2.2e-16	< 2.2e-16	0.0883	0.0658

Table 7. Results of the equivalence test of two structure indicators for the four-point variable for $x_3 = 2$ and $x_3 = 3$

		$X_3 = 2$		$X_3 = 3$	
		Standard sample	Modified sample	Standard sample	Modified sample
Population	□2	97.61	178.21	191.28	90.554
	p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

6. Conclusions

The aim of this study was to analyse the impact of the increase of the sample size in selected strata on the shape of sample variable distributions.

As can be seen from the completed simulation calculations, the non-proportional increase of the sample in strata produces significant differences between distributions of variables in the population and distributions obtained from individual tests. It is worth noting that distributions obtained for the standard sample are significantly different from distributions from the population. This is caused by rounding up to total minimum sample sizes in individual strata. An additional increase of the size in strata only deepens this effect. In the case of real research, another consequence of such plans will be an increase in the research costs. In view of the obtained results, it seems most advantageous to carry out random sampling strictly in accordance with the rules valid for the given sampling scheme. This is because even small changes may cause much bigger errors in the results than theoretically assumed.

Acknowledgements

Publication was financed from the funds granted to the Cracow University of Economics, within the framework of the subsidy for the maintenance of research potential.

References

Benadè, G., Gölz, P., Procaccia, A.D. (2019, June). No stratification without representation. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 281-314.

Cochran W. (1977). Sampling Techniques, 3rd edition, John Wiley & Sons, Inc.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, *70*(5), 646-675.

Formularz SOF1, Sprawozdanie z działalności fundacji, stowarzyszeń i podobnych organizacji społecznych (2018.10.11), Retrieved from: https://krakow.stat.gov.pl/osrodki/osrodek-badania-gospodarki-spolecznej-971/oso-formularze.

Peltoniemi, M., Heikkinen, J., Mäkipää, R. (2007). Stratification of regional sampling by model-predicted changes of carbon stocks in forested mineral soils.

Podgurski, A., Masri, W., McCleese, Y., Wolff, F.G., Yang, C. (1999). Estimation of software reliability by stratified sampling. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *8*(3), 263-283.

Sector Non-Profit in 2014, Statistical Analyses and Studies, Warsaw 2016, (2018.10.11). Retrieved from: https://stat.gov.pl/download/gfx/portalinformacyjny/pl/defaultaktualnosci/5490/1/5/1/sektor_non_profit_w_2014.pdf.

Steczkowski, J., Stefanów, P. (2009). Metoda reprezentacyjna w badaniu jakości wyrobów. Kontrola odbiorcza, Kraków: Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Zasępa, R. (1962). Badania statystyczne metodą reprezentacyjną. Warszawa: Państwowe Wydawnictwo Naukowe.

Zhao, X., Liang, J., Dang, C. (2019). A stratified sampling based clustering algorithm for large-scale data. *Knowledge-Based Systems*, *163*, 416-428.