ICFTE 6th International Conference on Future of Teaching and Education

Prague, Czech Republic 14 - 16 October 2022

The Use of Sampling in Quantitative Social Research

Dr. Tamar Tako Doreuli, Prof. Dr. Nino Durglishvili, Vano kechakmadze, Guguli magradze

Ivane javaxishvili tbilisi state university, Georgia

Abstract

Generalizing results based on sampling population research about the general population represents one of the most significant issues of quantitative social research. The models of relevantly formed sampling represent a necessary precondition for overcoming such a complicated and demanding challenge. The review of the sampling model and its appropriate weighing is provided in the article, based on which a quantitative sociological survey was conducted in 2020 in the capital of Georgia – Tbilisi, using face-to-face interviews. When formulating the sampling model, the following analytical tools were employed: stratification, cluster sampling, PPS methodology, systematic random sampling, probability weighting, and calibration. All full-aged citizens (18+) of Tbilisi have been specified as a target group; the database of population census has been used as a sampling frame, which encompasses all current households in Tbilisi. The sample size comprised 1,000 respondents. The two-phase, stratified-cluster model represented the type of sampling. Districts of Tbilisi presented strata, while cluster sampling was done by census-taking districts. The weighting model is constructed according to the sampling stages. The calculation of the standard error (5%) relied on the utilization of Thompson's formula (Thompson, 2012), and the design effect of cluster sampling (Kish, 1995). To grasp the conceptual essence of the design effect, the emphasis was placed on estimating distribution density for dependent observations (Kvatadze & Pharjiani, 2019). The central outcome of this research involves the representative sampling model of the total population, with a predefined standard error. However, it should be emphasized that such a model and individual techniques and approaches integrated within it can be successfully used in other similar research.

The process of sampling modeling underscored the significant importance of the theoretical sources and analytical tools used. In addition, exploring density estimation for dependent observations could prove advantageous for extending the research in relevant contexts.

Keywords: Quantitative research, stratified-cluster sampling, PPS, design effect, weighting

Introduction

Social science research is central in a "reality-based community." Social research helps us answer questions about the social world, raises new questions and may change how we look at the world as well (Neuman, 2014).

In quantitative social research, we have to take a sampling since we cannot study the entire population. The two main advantages of sampling are faster data collection and lower cost (Kish, 1995).

Generalizing results based on sampling population research about the general population represents one of the most significant issues of quantitative social research. The models of relevantly formed sampling represent a necessary precondition for overcoming such a complicated and demanding challenge.

Sampling is a critical component of most studies. It is described as the act, process, or technique of selecting a representative sample of a population to observe and analyze the characteristics of the entire population (Rahman et al., 2022). "The procedure by which the sample of units is selected from the population is called the sampling design" (Thompson, 2012).

Sample size calculation is a basic requisite for applied research in social science (Louangrath, 2019). The sampling technique and sample size determination are crucial in survey-based research problems in applied statistics (Singh & Masuku, 2014).

The sources related to the abovementioned issue are diverse. The PPS strategy has a practical advantage over other self-weighting procedures (Skinner, 2014). There exist numerous methods for selecting a fixed-size sample with probability proportional to size (PPS), with replacement (Abdulla et al., 2014).

Applications of PPS sampling in surveys can be split into two main categories: (a) sampling from area frames with a multi-stage design and (b) sampling ultimate units directly from a list frame. The classical situation for (a) is a multi-stage sample where primary sampling units are drawn with probability proportional to some measure of the unit's size (PPS). Typically, due to extensive stratification, just a few units are selected in each stratum (Ohlsson, 2000).

In quantitative research, sampling design is closely related to weighting techniques. Weights are commonly assigned to respondent records in a survey data file in order to make the weighted records represent the population of inference as closely as possible (Kaltan & Flores-Carvantes, 2003).

Methods

The two-phase, stratified-cluster sampling model represented the basic method. When using stratified sampling, the population is divided into subpopulations (strata). Taking a random sample from each subpopulation allows for controlling each stratum's relative size rather than letting random processes control it. A cluster is a unit that contains final sampling elements but can be treated temporarily as a sampling element itself.

In developing the sampling model, the subsequent analytical tools were employed: stratification, cluster sampling, PPS methodology, systematic random sampling, probability weighting, and calibration.

Results

A review of the sampling model and its appropriate weighing is provided in the article, based on which a quantitative sociological survey was conducted in 2020 in the capital of Georgia – Tbilisi, through face-to-face interviews.

The central outcome of this research involves the representative sampling model of the total population, with a predefined standard error. It should be emphasized that such a model and individual techniques integrated within it can be successfully used in other similar research.

The two-phase, stratified-cluster model represented the type of sampling. Strata were presented by districts of Tbilisi, while cluster sampling – by census-taking districts. The weighting model is constructed according to the sampling stages.

All full-aged citizens (18+) of Tbilisi have been specified as a target group; the database of population census has been used as a sampling frame, which encompasses all current households in Tbilisi. The location, address, and census-taking districts were provided in the database.

Strata were presented by districts of Tbilisi, while cluster sampling – by census-taking districts.

Description of the sampling stages:

The number of survey respondents was determined at the first stage of sampling. The sample size comprised 1000 respondents to reduce the margin of error.

The calculation of the standard error, set at 5% with a confidence level of 95% for proportional distribution, was grounded in the utilization of Thompson's formula (Thompson, 2012) and the cluster sampling's design effect (Kish, 1995)

$$\Delta^2 = 1.96^2 * \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 DEF_k \left(\frac{N}{n} - 1\right) \frac{S^2_{x,n,k}}{N_k}$$

In the context of this formula, the variables are defined as follows: N represents the total number of households in the population, N_k denotes the number of households in the k – th stratum, K signifies the number of stratums, and $S^2_{x,n,k}$ signifies the standard deviation of variable x within the k-th stratum, where n stands for the sample size of 1000.

In understanding the conceptual meaning of the design effect, there was a focus on estimating the distribution density in the case of dependent observations. In the case of dependent sampling, the kernel estimate of the density distribution is implemented through the conditionally independent sequence and chain-dependent observations (Kvatadze & Pharjiani, 2019).

The latter, in turn, is based on theorems on the asymptotic limit of standardized sums of sequences of conditionally independent and chain-dependent random vectors (Kvatadze & Shervashidze, 1987).

Proportional distribution was implemented at the second sampling stage as per Tbilisi districts. The number of survey respondents per district, comprised of five respondents, ensures a minor value of cluster sampling – "design effect."

Table 1

	Percentage Distribution	Number of Survey Respondents	Survey Respondents Per District	Number of Survey Districts
District 1 *(Mtatsminda)*	4.42%	45	5	9
District 2 *(Vake)*	9.17%	90	5	18
District 3 *(Saburtalo)*	12.47%	125	5	25
District 4 *(Krtsanisi)*	4.15%	40	5	8
District 5 *(Isani)*	11.36%	115	5	23
District 6 *(Samgori)*	16.17%	165	5	33
District 7 *(Chugureti)*	6.69%	65	5	13
District 8 *(Didube)*	6.53%	65	5	13
District 9 *(Nadzaladevi)*	13.58%	135	5	27
District 10 *(Gldani)*	15.46%	155	5	31
	100.00%	1000		200

In the third stage, district sampling was carried out per stratum from the sampling frame using systematic random sampling. The probability Proportional to Size (PPS) sampling method was used during the districts' sampling process.

The number of districts for sampling per stratum was specified according to Table 1.

In the fourth stage, starting point (one address) was sampled per district through systematic random sampling. The interviewer would start a survey from that address; the following target address would be the fifth from the step.

As for the address, selected at the fifth sampling stage – the respondents were sampled in the household by the nearest date of birth.

The weighting model depends on the stages of sampling. Weighting removes any bias that might result from having different kinds of people represented in the wrong proportion (Mercer et al., 2018). Correspondingly, the first stage of weighting was compatible with the district sampling. The following notations were used for the calculation of the probability of appearing on the list of district sampling:

$N -$ The number of households in the stratum;

N_i – The number of households in i – district;

$k -$ The number of sampling districts.

The probability of appearing on the list of district sampling is calculated according to the following formula:

$$P_i = k\frac{N_i}{N} \qquad (1)$$

While the second stage of weighting corresponded with household sampling per district. The currently used notations are as follows:

N_i – The number of households in i –district

n_i – The number of surveyed households in i – district

The probability of appearing on the list of the k - households in i – district is calculated according to the following formula (Lohr, 2021):

$$P(k|i) = \frac{n_i}{N_i} \qquad (2)$$

According to formulas (1) and (2), the calculation is as follows:

$$P_{i,k} = P_i \times P(k|i) = \left(k\frac{N_i}{N}\right) \times \left(\frac{n_i}{N_i}\right) = \frac{kn_i}{N} \Rightarrow Weight_i = \frac{N}{kn_i} \qquad (3)$$

The third stage of weighting complied with the respondent selection from the household, sampled from the district, where $L_{ik}$is the number of target members in (ik) – household.

The probability of the appearance of our preferable respondent on the sample list from

(ik) household is calculated according to the following formula:

$$P(l|ik) = \frac{1}{L_{ik}} \qquad (4)$$

The averaged characteristics have been established for the district at the following stage:

$$L_i = \sum_{k=1}^{n_i} L_{ik} \;\Rightarrow\; P(l|ik) = \frac{1}{L_{ik}} \approx \frac{n_i}{L_i} \;\Rightarrow\; P(l|ik) = \frac{1}{\bar{L}_\iota} \qquad (5)$$

A new formula is developed when using formula (4) and (5):

$$P_{ikl} = P_{ik} \times P(l|ik) = \left(\frac{kn_i}{N}\right) \times \left(\frac{1}{\bar{L}_\iota}\right) \Rightarrow WeightP_i = \frac{N}{kn_i} \times \bar{L}_\iota \qquad (6)$$

Based on this, the weights are formed for the respondents:

$$WeightP_i = \frac{N}{kn_i} \times \bar{L}_\iota \qquad (7)$$

The adjustment of age and gender was implemented as follows:

Gender represented two-group variables, 1 – a woman; 2 – a man.

The age group variable consisted of six categories:

1. 18-24
2. 25-34
3. 35-44
4. 45-54
5. 55-64
6. 65 +

Gender and age groups collectively created 6*2=12 groups called gender and age groups.

The number of population, according to the sampling frame, was notated as M_L in stratum, as $L -$ in gender and age groups and $n_{i,L}$ – the number of surveyed households in - i district.

The following formula is developed using formula (7):

$$\widehat{M}_L = \sum_{i=1}^{K} n_{i,L}\, WeightP_i n_{i,L} \neq M_L \Rightarrow Weit_{L,kor} = \frac{M_L}{\widehat{M}_L} \qquad (8)$$

According to formulas (7) and (8), the respondents' weight, by the districts in stratum, has been determined according to the following formula:

$$WeightP_{i,F} = WeightP_i \times Weit_{L,kor} \qquad (\mathbf{9})$$

Discussions

The process of sampling modeling underscored the significant importance of the theoretical sources and analytical tools used. Among them, "Survey Sampling" (Kish, 1995), "Sampling" (Thompson, 2012) and "Sampling: Design and Analysis" (Lohr, 2021) should be emphasized. Additionally, exploring density estimation for dependent observations could prove advantageous for extending the research in relevant contexts.

Conclusions

As a proper research has been conducted based on a sampling model we have considered and its relevant weighing, in which all significant target groups were presented with corresponding proportions for the implementation of a goal of research and objectives, we may conclude that such sampling model and all techniques used during its formation are effective and in certain cases, the researchers can use them successfully in other quantitative researches.

References:

Abdulla et al., (2014). On the Selection of Samples in Probability Proportional to Size Sampling: Cumulative Relative Frequency Method. *Mathematical Theory and Modeling, Vol.4* (6) 102-107. https://www.iiste.org

Kaltan, G., & Flores-Carvantes, I. (2003). Weighting Methods. *Journal of official Statistics, Vol.19* (2) 81-97. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/weighting-methods.pdf kish, L. (1995). *Survey Sampling.* New York: Wiley Classics Library

Kvatadze, Z., & Pharjiani, B. (2019). On the Exactness of Distribution Density Estimates Constructed by Some Classes of Dependent Observations. *Mathematics and Statistics. Vol. 7*(4), *135-145.* DOI:10.13189/ms.2019.070407

Kvatadze, Z., & Shervashidze, T. (1987). On Limit Theorems for Conditionally Independent Random Variable Controlled by a Finite Markov Chain. P*robability Theory and Mathematical Ststistic*s. *Lecture Notes in Mathematics. Vol. 1299, Springer-Verlag, (*250-258). DOI: https://doi.org/10.1007/BFb0078480

Mercer et al., (2018). *How different weighting methods work*. Pew research centre

Neuman, W. L. (2014). *Social Research Methods: Qualitative and Quantitative Approaches* (7th ed.). Essex: Pearson

Ohlsson, E. (2000). Coordination of PPS Samples Over Time. *The Second International Conference on Establishment Surveys. Alexandria VA: American Statistical Association*, (255-264). https://www.oecd.org/sdd/30890618.pdf

Rahman et al., (2022). Sampling Techniques (Probability) for Quantitative Social Science Researchers: A Conceptual Guidelines with Examples. *SEEU Review Vol.17*(1) 42-51. doi:https://sciendo.com/article/10.2478/seeur-2022-0023

Lohr, S. L. (2021). *Sampling: Design and Analysis*. 3rd Edition, New York: Wiley, https://doi.org/10.1201/9780429298899

Singh, A.S., & Masuku, M.B. (2014). Sampling Techniques and determination of sample size in applied statistics research: An overview, *International Journal of Economics, Commerce and Management Vol. 2(*11*), 11-22.* https://ijecm.co.uk

Skinner, C. J. (2014). Probability Proportional to size (PPS) Sampling. *Wiley Statsref: Statistcs Reference Online,* (1-5). *https://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat03346*

Thompson, S. K. (2012). *Sampling*. New York: Wiley