THE QUESTION OF REPRESENTATIVENESS OF THE SAMPLE SELECTED FROM THE SPECIFICALLY DEFINED POPULATION

Vladimíra Hovorková Valentová – Kateřina Gurinová

Abstract

This article is focused on recognition of various ways how to conduct sampling from the specifically defined population in order to hold the representativeness of the sample. Work of the authors results from the contract research of the Faculty of Economics TUL (Technical University of Liberec), No. of the contract TUL-00093961, called "The Survey of the Selected Indicators of the Traffic Services on the Tram Line No. 11". The aim of this survey was to count the number of ingoing and outgoing people in selected trams of the tram line No. 11 which goes from Liberec to Jablonec nad Nisou and back. The organization of such a survey was not the problem but the determination of which trams should be involved into the survey was. The limited financial recourses led to the organization of the sample survey. The main task was how to select the trams which should be the part of the sample in order to receive the representative sample. The authors of the paper present the final resolution here and they also discuss other possible ways of sampling included their advantages and disadvantages.

Key words: population, representativeness, sample, sampling, tram line

JEL Code: C18, C83

Introduction

The topic of this paper results from the contract research of the Faculty of Economics TUL (Technical University of Liberec), No. of the contract TUL-00093961, called "The Survey of the Selected Indicators of the Traffic Services on the Tram Line No. 11". The aim of this research was to find out the number of ingoing and outgoing passengers in selected trams of the tram line No. 11 going from Liberec to Jablonec nad Nisou and back. This data serves to the calculation of selected statistics of traffic services which are supposed to be a base for the traffic plan creation for the year 2015. Before we started to prepare the research, it had been clear that we should sample certain amount of units from the population. The reason was to

hold the costs on the acceptable level. Because of a great importance of the research results, the claim of the sample representativeness was laid. The principal question of this research was how many tram lines to select and how to select them.

1 Theoretical Base of the Research

The questions about the representativeness are closely connected with a certain type of sampling which is chosen in the certain situation. The sample is representative just when it is a small copy of the population and has as same characters as the population. The representativeness of a sample is a very often discussed topic in many scientific works, e. g. (Wilcox, Bellenger & Rigdon, 1994), (Kish, 1995), (Biemer & Lyberg, 2003), (Torbeck, 2012). Especially the sample representativeness is the matter of principle for obtained data, it has a great impact on the possibilities how to work with the data and what statistical methods can be used in the process of their analysis. Furthermore, it influences the possibility of data generalization.

We distinguish the purposive and the random sampling in relation to the way of the representativeness providing. But the usage of the purposive sampling cannot guarantee the possibility of generalization of the results coming from this survey sampling. Only the random sampling allows the correct application of the statistical inference methods and usage of all possibilities which they can provide.

The simple random sampling (SRS) is the most basic form of probability sampling. Its characteristic feature is that every unit which is a part of the population has the same chance to be selected. Its usage in practice is limited because researched problems often require the usage of some of more complicated forms of probability sampling. It is possible to use stratified sampling, cluster sampling or one-stage cluster sampling. Practical aspects of various types of random sampling were already described in (Gurinová & Hovorková Valentová, 2010).

The stratified sampling requires a division of the population into certain number of subpopulations which are called strata. It is necessary to divide the population with the help of a suitable criterion in order to receive the homogenous strata, i. e. the variability of units inside the subpopulations should be very low. In the next step a certain amount of units is selected e. g. taking a simple random sample. The question of stratification, especially homogeneity and the number of strata is often discussed in scientific books, e. g. (Stephan,

1941), (Kish, 1995), (Cochran, 1977), (Triola, 2010). It is apparent that the choice of a stratification criterion is crucial for optimal application of stratified sampling.

Cluster sampling is generally carried out in more steps but the most common form is two-stage cluster sampling. We suppose division of the population into some subgroups (clusters) which are also called primary sampling units. In the first step we select certain amount of primary sampling units (psus) and from these selected psus some secondary sampling units (ssus) are selected. This type of sampling is more difficult than the stratified sampling in terms of using mathematical-statistical methods and requires a proper arrangement. The questions about the construction of the two-stage cluster samples and problems connected with it are mentioned in works of many authors, e. g. (Groves, Fowler, Jr., Couper, Lepkowski, Singer & Tourangeau, 2009), (Kish, 1995), (Cochran, 1977). The practical aspects of two-stage cluster sampling application are also contained in (Gurinová & Hovorková Valentová, 2011 and 2013).

The special form of two-stage cluster sampling is one-stage cluster sampling. In the first step we select a few psus, less than in case of two-stage cluster sampling. These psus are analysed in the second step as they are, i. e. all elements of these clusters are observed. This type of sampling is used in situations when the population contains great amount of elements and the statistical units are spread in a large area. We found the comparison of the stratified sample and the one-stage sample to be interesting and it is introduced in (Lohr, 2010).

The sample size determination is a very important factor which influences every survey sampling. It depends on many requirements and presumptions which can often contradict each other. From the statistical point of view it is necessary to require the sufficient estimates reliability and the important thing is also to determine the maximum standard error of estimation. Many authors write about this topic in their works, e. g. (Kish, 1995), (Biemer & Lyberg, 2003), (Kadam & Bhalerao, 2010), (Triola, 2010). The maximum survey costs are often the limiting factor which is usually in conflict with the requirements mentioned above. The time factor is another important indicator which affects the decision about arranged sample survey. It means time requirements for arrangement and organization of a survey, following data analysis and formulation of outcomes.

2 The Creation of the Tram Lines Selection Methodology

Before thinking about the suitable type of sampling particular tram lines, it was necessary to determine the sample size. As was mentioned in the previous chapter, when we determine the

sample size we have to take into account the estimates reliability, the standard error of estimates and also the survey costs. Sometimes it is difficult to combine all these requirements. Here we determined the sample size firstly in the term of survey costs. Firstly, we count the maximum of observed lines when we suppose work of 4 persons in one day if these persons go on the tram line No. 11 from Liberec to Jablonec nad Nisou and back without any break. We found out that it would be possible to observe 28 lines in working days. The sum of all lines going on that way in working days is 71. The ratio of observed lines is then 39.4 per cent which is the sufficient number. Then, we select the similar proportion of lines at the weekend.

While determining the way of the tram lines selection it is necessary to define the population and how many units it contains. In this case we did not work with just one population but with three ones. The first population is made with all the lines going in working days. Their sum is 71. The second population contained all the lines going on Saturday and the third population was made with all the lines going on Sunday. The reason for division of lines into three populations was the different timetable and the different number of going lines (59 on Saturday, 55 on Sunday).

Let us focus on the lines selection in working days. From the first while we did not think of the observation of the lines of all the working days. The reason was the effort to achieve as low costs as possible. Therefore, we selected just three working days – Monday, Wednesday and Friday. The selection of Monday and Friday was based on the idea that these days have specific characteristics in comparison with other working days. On Mondays a lot of people using the public transport come back to schools, work, school hostels, campus etc. Therefore, we can expect increased interest in the public transport, especially in the morning. The situation on Fridays is opposite in comparison with Mondays. Many people leave the school hostels, dormitories, campus and schools for home or the Liberec inhabitants leave the city for the weekend. Wednesday was selected as the "common" working day, i. e. we suppose the similar interest in the public transport on Tuesday and Thursday as well.

When we were making the decision about the way of the lines selection, we thought of the simplest way first – simple random sampling. We took a few simple random samples from the population and it showed that it is not the best way how to make an idea about the lines occupancy uniformly during all the day. We found out that in some cases even three lines in one hour were included into a sample while some hours stayed omitted. Then, we made an idea that each hour of the day should have its own representative in the sample. All following plans how to carry out the sampling resulted from this idea then.

We could exclude one-stage cluster sampling from the offer of possible ways because it would be in the conflict with the idea mentioned above. We also did not find adequate reasons for a choice of cluster sampling. So, we focused on stratified sampling. All the population was divided into 21 strata. The stratum is defined as the group of lines leaving the Rybníček station (in the direction from Liberec to Jablonec nad Nisou) in every hour of the day. The reason why the Rybníček station was the starting station was that some of lines do not go from the first station on the way which is the Viadukt station. But all the lines go from or through the Rybníček station. Briefly, the population is made by all the lines going from or through the Rybníček station. The determination of suitable subgroups whose inner variability of the observed variable is low is the important requirement in the process of stratified sampling. The observed variable was defined as the number of ingoing and outgoing passengers. It is possible to suppose that the variability of the number of ingoing and outgoing passengers would be in the same hour similar, therefore we can estimate that the strata were defined in the suitable way.

Stratified sampling was carried out in two steps. In the first step one line in each hour was randomly selected. Let us add the information that in some hours just one line goes. So, this line was selected (i.e. the stratum contained one unit only). Twenty one lines was selected by this way in working days, twenty on Saturday and twenty one in Sunday. The rest of lines in working days was selected in the second step. The file from which other lines were selected was made by all the lines which have not been selected yet. Seven lines were selected in the second step and we carried out it as simple random sampling. Let us make a note that the line going at 3:55 a.m. from the Rybníček station was excluded from the population. The reason was purely practical – observers´ disability to come to the station at that time. We carried out the stratified sampling without replacement because sampling with replacement is not meaningful here – we need not to observe the line which has already been observed. The selection of lines in the direction from Jablonec nad Nisou to Liberec was carried out in the context of the first selection – we selected the lines which leave Jablonec nad Nisou, Tyršovy sady immediately after coming the randomly selected line from Liberec to Jablonec n. N. The randomness of the sample of lines going from Liberec to Jablonec n. N. would provide the representativeness of the sample of lines going in the return direction.

3 Discussion

If we discuss the conducted sampling, we will certainly find a space for a polemic about the way of the lines selection. While resuming advantages and disadvantages of random sampling mentioned in the previous chapters we find again that the choice of stratified sampling was the only possible way if we wanted to receive the sample which can be a base for the creation of the traffic plan for next years. The use of simple random sampling is not able to guarantee uniform representation of every hour of the day in the sample which is an important fact for the traffic plan creation. One-stage cluster sampling would mean that we select some hours only and others not and then we would observe all the lines in selected hours. Furthermore, the requirement of inner stratum heterogeneity would not be performed. In this time we do not find any possibility how to define the groups to achieve the requirement performance. Cluster sampling would require division of the population into several primary sampling units. Moreover, it is suitable in cases when we can suppose placement of the statistical units on a large area. This problem did not arise here so, we do not have the reason to use cluster sampling.

If we made a conclusion that stratified sampling is the best way how to select the lines again, we can discuss about the definition of the strata. We also had an idea to define the strata as the collection of lines going in morning rush hours, afternoon rush hours etc. But we met two problems – firstly, it is difficult to determine the borders of rush hours, and secondly the danger that the sample will not contain the representative of each hour of the day arises again.

Conclusion

As has been mentioned in the previous text the guaranty of sample representativeness is sufficient number of elements random sampling. We can carry out the selection of determined number of lines by several ways. Firstly, it is necessary to decide what the sample will look like in order to be possible to use it for the determined purpose. The use of SRS is very easy and the statistics describing features of the sample are simply countable. It seems to be a good choice. The truth is that taking a simple random sample does not allow us to receive information about the utilization of trams capacity in every hour of the day which is the important fact in the process of the traffic plan creation. Then, we should think about more complicated form of probability sampling. So, it is not meaningful to try to avoid complications which can arise when we use some of more complicated forms of random sampling. The most important thing is that the sample should be a good base for making decisions of a great importance.

Acknowledgment

This paper was written in the frame of work on the contract research of the Faculty of Economics TUL (Technical University of Liberec), No. of the contract TUL-00093961, called "The Survey of the Selected Indicators of the Traffic Services on the Tram Line No. 11".

References

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. New Jersey: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

Groves, R. M., Fowler, Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. (2nd ed.). New Jersey: John Wiley & Sons.

Gurinová, K., & Hovorková Valentová, V. (2011). *Advantages of Two-Stage Cluster Sampling when Carrying out the Random Sampling from the Population of the Czech Republic*. In Kocourek Aleš (Ed.), *Proceedings of the 10th International Conference Liberec Economic Forum 2011* (pp. 139-146). Liberec: Technical University of Liberec.

Gurinová, K., & Hovorková Valentová, V. (2010). *Možnosti provedení náhodných výběrů z populace ČR za účelem zkoumání vývoje hospodářských ukazatelů*. In Moc Radek (Ed.), *FernStat_CZ 2010, VII. ročník mezinárodní konference, Ústí nad Labem, 23.-24. září 2010. Sborník příspěvků* (pp. 35-42). Ústí nad Labem: UJEP, FSE.

Gurinová, K., & Hovorková Valentová, V. (2013). *Praktická aplikace dvoustupňových náhodných výběrů v souboru obcí ČR*. In Jedlička Pavel (Ed.), *Hradecké ekonomické dny 2013, mezinárodní vědecká konference, Ekonomický rozvoj a management regionů, Hradec Králové 19. a 20. února 2013, sborník recenzovaných příspěvků* (pp. 154-161). Hradec Králové: Nakladatelství Gaudeamus, Univerzita Hradec Králové.

Kadam, P., & Bhalerao, S. (2010). Sample Size Calculation. *International Journal of Ayurveda Research, 1*(1), 55-57. Retrieved from http://search.proquest.com/docview/866307618?accountid=17116

Kish, L. (1995). *Survey Sampling*. New York: John Wiley & Sons.

Stephan, F. F. (1941). Stratification in Representative Sampling. *Journal of Marketing (pre-1986)*, *6*(1), 38-46. Retrieved from http://search.proquest.com/docview/209304860?accountid=17116

Torbeck, L. D. (2012). Representative Sampling. *Pharmaceutical Technology*, *36*(4), 38-40. Retrieved from http://search.proquest.com/docview/1015033611?accountid=17116

Triola, M. F. (2012). *Elementary Statistics Technology Update* (11th ed.). Boston: Pearson Education.

Wilcox, J. B., Bellenger, D. N., & Rigdon, E. E. (1994). Assessing Sample Representativeness in Industrial Surveys. *The Journal of Business & Industrial Marketing*, *9*(2), 51-61. Retrieved from http://search.proquest.com/docview/222004694?accountid=17116

Contact

Vladimíra Hovorková Valentová
Technical University of Liberec
Studentská 2, Liberec
vladimira.valentova@tul.cz

Kateřina Gurinová
Technical University of Liberec
Studentská 2, Liberec
katerina.gurinova@tul.cz