APPLICATION OF THE METHOD OF SAMPLES ON THE EXAMPLE OF THE POPULATION OF VALUE PURCHASES IN THE ONLINE SHOP *SUPERSTORE SALES*

Kristina Zogović
University of Belgrade, Faculty of Economics, Serbia zogovic.kristina@gmail.com

Abstract: To collect information about mass phenomena, research is often used in practice through samples. The subject of statistical research is the mass phenomenon of variable character (population), about which, with appropriate statistical methods, we make conclusions. A key precondition for valid validation is the representativeness of the sample. This assumption is fulfilled if the sample structure corresponds to the population, that is, representativeness can be achieved by the correct selection of the elements of the set. Based on the selected elements, using statistical tests, we draw conclusions about the character traits of the sample, and further on the basis of the sample on the characteristics of the basic set. Using statistical methods, we investigate the reliability of the estimation and the accuracy of the estimation. These procedures are referred to as the sampling method. The aim of this research is to examine the accuracy of the grades obtained with the help of a simple random method, a stratified sample, and a systematic sample. The sample plan represents the method of selecting elements from the population and forming a sample, while the sample is a set of selected elements of the basic set. In carrying out the statistical research, it is necessary to define a precise sample selection plan in advance, in which way we select the elements and what is the criterion for determining the elements of the sample. The sampling plan consists of several chronological steps that form an integral whole: defining the population, identifying the sample frame, selecting sample types, defining the sample size, and implementing the process of selecting elements as the phase of implementation of the plan. The dominant criterion for sample distribution is probability, i.e. When choosing elements in a sample, each element has a predetermined probability of choice, with probabilities different from zero. Of the probability-based samples, the most commonly used are: a simple random sample, a stratified sample, and a systematic pattern. In this paper we analyzed surveys through a sample population of 5,000 observations of the online store Superstore sales for the period from 2009 to 2012. The observed mark is the value of the purchased purchases. Based on the data on the completed purchases, it is possible to obtain a number of very useful information about customer preferences in terms of: product categories, product dimensions, ordering time, product values and a number of cross-related data, such as the value and type of purchases by individual provinces and regions, or the value of purchases by the type of delivery and similar. By comparing the obtained results from the above samples we can conclude that the most precise is stratified sample.

Keywords: sampling, assessment and accuracy of the assessment, sampling methods.

1. INTRODUCTION

Gathering information about a mass phenomenon can be based on a population survey, which requires organized research preparation, longer research time and very high costs. This type of statistical research is called a census. The analyzed phenomenon that we want to make conclusions and find out its characteristics it is called a statistical gathering, a basic set or a population. Although the census itself gives us full and accurate information about the characteristics of the population, given the listed negative characteristics of the census, very often in practice, the research is carried out through sample survey. There is no method of grading on the sample survey as the universally best method, and we can not rely on one unique method to provide us with completely reliable and precise conclusions. The choice of the sample depends on the observed phenomenon, on the manner of data collection, on the characteristics of the set we want to explore, from the availability of resources and other factors. The aim of the survey through the sample is to conclude on the characteristics of the population based on the selected population elements, that is, the observation units. The size of the sample depends on the requirements of the applied tests, the size of the population, the applied methodology, and similar. Irrespective of the specific size of the sample, it must be sufficient to derive relevant and valid statistical conclusions from it. The idea of this paperwork is, through practical work, to demonstrate and confirm the higher accuracy of stratified in relation to a systematic and free random sample.

2. DESCRIPTION OF POPULATION

In the analysis of the sample survey, a population of 5,000 observation orders of the online store Superstore sales for the period 2009-2012 was observed. The observed mark is the value of the purchased purchases. In this period, the delivery was done in three ways: Delivery Truck, Regular Air and Express Air (fast ones). Depending on the buyer and the value of the purchase, we distinguish five priority groups (critical, high, medium, low and unspecified group) and sixteen types of rebates. This store keeps detailed records of each customer (attributes are: name, province and region of the buyer, branch from which it comes) and to each store (category and subcategory of purchased product, product name, type and dimensions of packaging, date of delivery) which greatly facilitates business analysis at any time. Based on the data on completed purchases, it is possible to obtain a number of very useful information about customer preferences in terms of: product categories, product dimensions, ordering time, product values and a number of cross-related data, such as the value and type of purchase by individual provinces and regions, or the value of purchases by the type of delivery and similar. Below we will show the results obtained for a random sample, obtained by using the RANDBETWEEN function (each element of the basic set has the same probability to be selected and the choice is non-returnable).

3. SAMPLE COMPENSATION

A simple random sample

This type of sample has some theoretical advantages, but at the same time there are practical shortcomings when it comes to its application in statistical surveys. A simple random sample is obtained from a population of a certain number of elements if the selection is performed so that each sample selected from the population has the same probability of choice. Only that simple random sample chosen with strict adherence to all theoretical assumptions has a sufficient level of reliability. On the basis of this, we can quantify the degree of reliability and evaluate the parameter. However, the fulfillment of all assumptions requires investments that are very often higher than the effect of the realization of a particular research. In Canada, Superstore sales online store is registered. In the period from 2009 to 2012, 5,000 purchases were registered. Data for 250 (n) purchases was collected to evaluate the mid-size value of purchases. In order to select 250 units of the basic set, we first number all units with numbers from 1 to N, i.e. from 1 to 5,000. As in practice, the correction of the final population is ignored if the population fraction (n / N) does not exceed 5%, and often even 10%, that is, the number n[82] is obtained as 5% of 5,000. The factor (N-n) / N, which can also be written as

1- (n / N), is called the correction factor for the final population. If the population is large in relation to the size of the sample, so that the fraction of the sample n / N is small, this factor becomes familiar to the unit[83]. With the RANDBETWEEN function in Excel, 250 numbers from 1 to 5000 were selected randomly. In this way we get the following random sample:

Table 1. A free random sample of buying value (sorted in decreasing order) in Superstore sales store

Random number	Ordinary number in a sample of 250 elements	Customer name	Value of purchase
3904	1	Emily Phan	89.061,05
1038	2	Jasper Cacioppo	45.923,76
3532	3	Craig Carreira	41.343,21
4589	4	Dennis Kane	33.367,85
...	...	...	...
4671	4994	Carlos Soltero	4,99
4902	4995	Don Weiss	4,94
4425	4996	Jeremy Farry	3,96
3962	4999	Ken Dana	3,20
4416	5000	Ricardo Emerson	2,24

The average purchase value is: $\bar{Y} = \frac{1}{5000}\sum_{i=1}^{5000}$ $\frac{1}{5000}\sum_{i=1}^{5000}$ y_i = 1.818,90$

[82] Lj.Petrović (2013): *Teorija uzoraka i planiranje eksperimenata*, Ekonomski fakultet, CID, Beograd, 2015, str 20

[83] Lj. Petrović (2012): *Zbirka rešenih zadataka iz teorije uzoraka i planiranje eksperimenata*, Ekonomski fakultet CID, Beograd, str 6

If the variance of a simple random sample variance is:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 14.979.048,48$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 14.979.048,48$$

the mean score variance is:

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)\hat{V}(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = 56.920,38$$

The results obtained by the stratified sample are as follows. The sample is divided into four strata. As in the previous section, the random sample was selected using the RANDBETWEEN function.

Stratified sample

A stratified sample is applied for those populations where there is a high degree of heterogeneity, that is, in the case of marked variability of the trait. Bearing in mind the pronounced variability and limitations in terms of sample size, it is difficult to obtain estimates of a satisfactory level of precision based on a simple random sample. By stratification, we achieve a significant improvement in the accuracy of the random sample.

This type of sample is based on the division of the population into strata, and from each strata simple random samples of a smaller scale are selected. The number of sample units from the strata is determined on the basis of proportionality or based on the assessment of the significance of each individual strata. The initial step in the stratification of the sample is to edit the strata in order to distinguish one from the other as much as possible and that the elements within each strata are as homogeneous as possible. In addition, it is possible to execute previously if we have information about the structure of the observed set. Information refers to the features that are the subject of the research and whose characteristics are assessed or tested. When assigning a total sample to individual strata we can apply a proportional and optimal schedule. If the proportions are retained in the strata as well as in the population ,the answer is the proportional schedule.

In our example of buying value from each strata, by using a proportional allocation method, we choose simple random samples. The size of the simple random samples that are drawn from the stratum in relation to the size of the stratified sample is equal to the proportion of the corresponding stratum in the basic set. Two analyzes were performed with different groupings (intervals) and in both cases IV strata were formed. In the case of a stratified sample an interval was applied based on which results were obtained which are shown in the table below and obtained the following cross-sections:

Table 2. Cross sections in the stratified sample Stratum

	Stratum	**Size of the stratum**
First cross sections	56,57	400
Second cross sections	113,15	1058
Third cross sections	169,72	1158
		2384
Totall:		5000

The last value in the column in the table bellow, the cumulative frequency (226,30) is divided by the number of strata (4) and the first cross section value is obtained: 226,30 / 4 = 56,57. The second section is obtained as a double value of the first cross section (56.57 * 3 = 113.15), and the third section is analogous (56.57 * 4 = 169.72).

Table 3. Intervals with a stratified sample

Purchase value	**Summed purchase value intervals**	**Frequency f**	**Root of f**	**Cumulative frequency**
od 80.001,00 do 90.000,00	od 42.001,00 do 90.000,00	1	1,00	1,00
od 72.001,00 do 80.000,00	od 30.001,00 do 42.000,00	3	1,73	2,73
od 66.001,00 do 72.000,00	od 24.001,00 do 30.000,00	16	4,00	6,73
od 60.001,00 do 66.000,00	od 18.001,00 do 24.000,00	35	5,92	12,65
od 54.001,00 do 60.000,00	od 12.001,00 do 18.000,00	80	8,94	21,59
od 48.001,00 do 54.000,00	od 6.001,00 do 12.000,00	265	16,28	37,87
od 42.001,00 do 48.000,00	od 4001,00 do 6.000,00	247	15,72	53,59
od 36.001,00 do 42.000,00	od 2001,00 do 4.000,00	481	21,93	75,52
od 30.001,00 do 36.000,00	od 1.401,00 do 2.000,00	330	18,17	93,69
od 24.001,00 do 30.000,00	od 701,00 do 1.400,00	640	25,30	118,98
od 18.001,00 do 24.000,00	od 401,00 do 700,00	518	22,76	141,74
od 12.001,00 do 18.000,00	od 201,00 do 400,00	779	27,91	169,65
od 6.001 do 12.000,00	od 101,00 do 200,00	770	27,75	197,40
od 0,00 do 6.000,00	od 0,00 do 100,00	835	28,90	226,30

In the Cumulative Root column f the three most favorable values were obtained for the obtained cross section values (56.57-53.59; 113.15-118.98; 169.72-169.65), and the fourth value was taken from the cumulative value of the frequencies (the last Value from the column, i.e., 226,30). The given values are defined by confidence intervals. Based on the values obtained by interval and finding the three sections, stratums, their size, as well as the representation in the sample, which can be seen in the table below (Table 4).

Table 4. Stratified sample

Stratums of purchase	Cumulative root of f f	Size of startums N_h	Representation in a sample of 250	Mean of stratums $\bar{y}_h$
6.000,00 - 90.000,00	53,59	400	20	11.649,95
1.400,00 - 6.000,00	118,98	1058	53	2.963,10
400,00 - 1.400,00	169,65	1158	58	796,16
0,00 - 400,00	226,30	2384	119	158,40
Total		5000	250	

The average purchase value is: $\bar{y} = \frac{\sum N_h \bar{y} h}{N} = 1.818,90$

$$\bar{y} = \frac{\sum N_h \bar{y} h}{N} = 1.818,90$$

Stratum variance estimates are: $s_1^2 = \frac{1}{n_1 - 1} \sum [(y_{1i} - \overline{y_1})^2$

$$s_1^2 = \frac{1}{n_1 - 1} \sum [(y_{1i} - \overline{y_1})^2$$

Table 5. Stratums and variance estimation

Stratums of purchases	Estimates variances of stratums
od 6.000,00 do 90.000,00	22.542.959,26
od 1.400,00 do 6.000,00	24.266.879,15
od 400,00 do 1.400,00	312.726,50
od 0,00 do 400,00	27.696,36

The assessment of the variance of the obtained environmental assessment is:

$$\hat{V}(\bar{y}) = \frac{1}{N^2}\sum N_h\,(N_h - n_h)\frac{s_h^2}{n_h} = 14.333{,}68$$

$$\hat{V}(\bar{y}) = \frac{1}{N^2}\sum N_h\,(N_h - n_h)\frac{s_h^2}{n_h} = 14.333{,}68$$

If we compare the obtained variance estimates of the stratified sample (14.333,68) and variance of a simple random sample (56.920,38), we conclude that a stratified random sample is **more accurate** than a simple random sample. Below we can see an example of a systematic random sample and compare it with results obtained with a stratified sample and a simple random sample.

Systematic random sample

The number i = 3 was chosen randomly in the systematic sample preparation, and the value for k was determined using: k = 5,000 / 250 = 20. In a systematic pattern of correlation within the strata, they are not always positive. For example, in the first systematic pattern, 7 of the 13 elements are larger than the center of the strata they belong to (ie deviations are positive), while the other six are not. There is no systematic pattern without positive deviation. This fact results in different precision of the systematic and stratified sample. The sampling environment variance is:

$$s^2 = V(\bar{Y}_{sy}) = \frac{1}{n^2 k}\frac{1}{n^2 k}\ [\ \sum_{i=1}^{k}\ \sum_{j=1}^{n}\ \Sigma_{\downarrow}(i=1)^{\uparrow}k\equiv(\Sigma_{\downarrow}(i=1)^{\uparrow}n\equiv\ y_{ij})^2\ -\ \frac{1}{k}$$

$$\frac{1}{k}(\sum_{i=1}^{k}\ \sum_{j=1}^{n}\ \sum_{i=1}^{k}\square\ \sum_{i=1}^{n}\square\ y_{ij})^2] = 2.116{,}17$$

The list of population units based on which the sample is selected is randomly arranged, so a systematic sample can be identified with a simple random sample without returning and apply the same evaluation formula of variances[84]. Sampling variance estimates it is:

$$\hat{V}(\bar{y}_{sy}) = \frac{s^2}{n}\frac{s^2}{n}\ (1\ -\ \frac{n}{N}\frac{n}{N}\ ) = \frac{2.116{,}17^2}{250}\frac{2.116{,}17^2}{250}\ (1\ -\ \frac{250}{5000}\frac{250}{5000}\ )\ = \mathbf{17.017{,}10}$$

By comparing the obtained results from the above samples, we can conclude that the most precise stratified sample

4. CONCLUSION

The method of research based on samples has certain characteristics, that is, positive and negative sides, which we must be aware of when conducting the research itself. The advantages of this method are: greater efficiency, actually,the speed of data collection and processing, the higher the speed of obtaining the results, the significantly lower research costs compared to the implementation of the census, the higher the reliability of the results, because the research is dominated by professional staff and specially prepared interviewers whose engagement is justified due to the reduced volume of activities, the greater flexibility in the different types of data that can be collected, it is possible to optimize the sample size with the appropriate (acceptable) level of risk. The disadvantages of this method are: the results contain a sample error, special training of staff involved in the research and project management by the statisticians is required; the sample does not provide data for each unit of the observed population; in the intentional samples, the estimation of the population parameters can not be carried out in strictly scientific terms . Based on the above sampling methods and comparing the obtained results from the above samples, we can conclude that the most accurate stratified sample with a rating is $ 14,333.68; after him it is a systematic sample, with an estimate of $ 17,017.10 and the most imprecise is a free random sample with an estimated $ 56,920.38. As the theory claims, the most accurate grades are obtained by the stratified sample method. The mean score of both stratified

[84] Opus cit.br 5, str 75

samples is $ 1,818.90, which is identical to the average median in the observed population. Based on this, we can conclude that the average size of the purchase value in the online store Superstore sales is very high.

LITERATURA

[1] Dejvid, A., Kumar, B., Dej, DZ., (2008), *Marketinško istraživanje*, Ekonomski fakultet, Beograd

[2] Lindgren, M., Bandhold, M.,(2003), *Scenario Planing: The link between future and strategy,* Palgrave MacMillan

[3] Lj. Petrović (2015): *Teorija uzoraka i planiranje eksperimenata*, Ekonomski fakultet, CID, Beograd

[4] Lj. Petrović (2015): *Zbirka rešenih zadataka iz teorije uzoraka i planiranje eksperimenata*, Ekonomski fakultet CID, Beograd

[5] Lj. Petrović, (2015): *Teorijska statistika. Teorija statističkog zaključivanja*, Ekonomski fakultet, CID, Beograd

[6] Mann S. (2009): *Uvod u statistiku,* Ekonomski fakultet CID, Beograd

[7] Singelton R.A., Jr., Straits,B.C.,&Straits,M.M. (1999) *Approches to Social Researches* (3rd ed.). New York:Oxford Univerity Press

[8] Badi H. Baltagi, Sadka E., (2008), *Measuring marketing power: Contributions to Economic Analisis, Emerald Group Publishing Limited*

[9] Stapenhurst, T., (2009), *The Banchmarking book: A how-to-guide to best practice for managers and practiotioners*, Butterwort-Heinemann

[10] Žižić M., Lovrić M., Pavličić D., (2006): Metodi statistike analize, Ekonomski Fakultet CID, Beograd

[11] Yin, R.K. (2009) *Case Study Research design and methodes*, (4rd ed.). Nashwille: Peabody Kollege, Vanderbilt University