WAYS OF TAKING RANDOM SAMPLES FROM A POPULATION FOR THE NEEDS OF AN ECONOMIC INDICATORS ANALYSIS

Kateřina Gurinová
***Vladimíra Hovorková Valentová**

Technical University of Liberec
Faculty of Economics
Studentská 2, 461 17, Liberec 1, Czech Republic
katerina.gurinova@tul.cz
* Technical University of Liberec
Faculty of Economics
Studentská 2, 461 17, Liberec 1, Czech Republic
vladimira.valentova@email.cz

Abstract

The aim of this article is to provide a list of possibilities how to take random samples from the Czech Republic's population and, consequently, their comparison. The comparison of the presented methods was carried out with the help of selected statistic, which enables us to choose a method that brings in the most accurate estimates. However, we must not forget the fact that when taking random samples in practise, we are limited by financial means, working force that participates in the survey, and other factors. We are, therefore, presenting an ideal solution in the first example and an optimal one in the second example while having taken into account all influencing factors.

Introduction

This article was elaborated with the financial help of the project no. 1101 of the Fund for the Development of Higher Education Institutions called Creation of a new module "Statistical Data Analysis of Questionnaires" and in connection with a project registered as WD-30-07-1 in the research programme of the Ministry for Regional Development. This project called "Innovation Approach to Analysis of Disparities on Regional Level" has been carried out at the Faculty of Economics, Technical University of Liberec during the years 2007- 2011. The question how to take the best random samples from a population arose while solving tasks of the above mentioned project, and during consultations with students studying at the Faculty of Economics, Technical University of Liberec (EF TUL). We have been recently working, within the frame of the project WD-30-07-01, with the population of the Czech Republic's municipalities, where values of certain economic indicators were being elicited. The aim of this article is to provide other possible ways of taking random samples of a certain number of municipalities in the Czech Republic, which will be instrumental in exploring selected economic indicators. This article is a follow-up to the research results published in [1]. New suggestions for further research were received after publishing the article [1], and these are presented in this article.

1 Theoretical look at a random sample

Let us quickly recapitulate how a random sample is defined, and what kinds of probability samples we have. A sample is called random when data are obtained by random sampling.

Randomness ensures the representativness of a sample and statistics obtained by such a sample can be generalized to a population by the methods of mathematical statistics. From the probability point of view, random sampling can be implemented by equal or unequal probabilities.

The simplest kind of a random sample is a simple random sample - SRS. It is a direct selection of elements from an unsorted population. Each element, which is in the population during this draw, has, during each draw, the same probability of being drawn. We distinguish between two simple random samples, namely a simple random sample with replacement and a simple random sample without replacement. A sample with replacement has a pattern of independent trials. The probability that each element will be chosen is the same for all draws (1/N), and the size of the population does not vary during the draw. A sample without replacement has a pattern of dependent trials, the probability that each element will be chosen rises with each draw. The size of the population decreases with each following draw. However, SRS is often unsuitable due to its simplicity. Therefore, other more complex kinds of random samples, which enable us to see the complicated reality better, are used.

One of the more complicated kinds of sampling is stratified sampling. The population must be firstly subdivided into groups or by other name strata. A correct allocation of the population into strata has a great impact on the sample quality. Strata are defined as groups of elements, which are somehow similar; it means that the strata are more homogenous inside than the sample as a whole. As S. L. Lohr says in [5], stratification is the most effective when means in strata vary a lot. These groups of elements can be both natural and artificial. A random sample of a given number of elements is taken in each stratum. The most common sample is proportional allocation, where sample sizes in each stratum are in due proportion to the sizes of the strata. Yet, a different approach can be taken, such as taking the same, previously stated, number of elements from each stratum. In this case we call it uniform sampling.

Another option is optimal allocation of sampling into strata, it means that sample sizes are not only proportional to the sizes of strata, but furthermore, their variability is also taken into account. The drawback of this procedure is its relative complexness. Stratified sampling is complex as it requires certain preliminary information necessary for assigning elements into strata. The next drawback of stratified sampling is the fact that is leads to a relatively large space variance. All in all, it is more demanding survey organisation and data processing wise, which of course increases the survey expenses. The advantage, in comparison to SRS, is the fact that stratified sampling increases efficiency of estimators.

Cluster sampling is considered as a more complicated kind of a random sample. Its simplest type is two-stage cluster sampling, yet, the procedure can be generalized into more stages. A population must be divided into groups. Groups of units, called primary units, are randomly taken from the population during the first stage. Then, during the second stage, statistic units, called secondary units, are randomly selected from the primary units. The advantage of such a kind of sampling, compared to stratified sampling, is the fact that space variance of the selected units is significantly smaller, which leads to a reduction of the survey costs. The disadvantage is that it brings in less reliable reasons within the same sample size than the simple random sampling or stratified sampling. It is due to the fact that some primary units are entirely left out during this procedure; therefore, there is no information about them available. Cluster sampling requires very precise preparation, and its processing by mathematical-statistical methods is more complex.

2 Summary of obtained results

The aim of the previous activities was to apply different kinds of random samples to specific data and to carry out a comparison of the obtained samples in terms of their representativness. The population we worked with consisted of 6,248 Czech municipalities. The researched economic indicator was unemployment rate in %, in the year 2006. The Czech Statistical Office (CSO) supplied the data about this indicator in all Czech municipalities. The sample size was estimated as 520 units. Such a size is big enough to allow us to generalize the results, and, in addition, it allowed us to carry out systematic sampling. We took 30 random samples from the given population. They represented 10 SRSs, systematic sampling was used in 5 cases, and other 5 samples were obtained by a random number generator, which was run in the statistic software STATGRAPHICS CENTURION XVI. Other 15 samples were obtained by stratified sampling (uniform, proportional, and optimal allocation); the last 5 samples were taken by two-stage sampling. We used several criteria to compare the quality of our estimates obtained by different kinds of random samples:

- Standard error of the mean
- Mean deviation
- Relative gains from stratification

It is well known that when a real value is replaced by an estimate obtained by sampling, so called sampling error occurs . It is impossible to define it in a real situation; we can only speculate about some of its allocation characteristics. Therefore, measures of statistic variation t are used to measure the quality of an estimate, the most common is the mean squared error - page 28, in [4]:

$$E\left[(t-\Theta)^2\right] = D(t) + b_t^2(\Theta). \qquad (1)$$

This measurement measures the error of point estimation. In the case that the sample characteristic is an unbiased estimator of the population characteristic, the mean square error equals variance. A standard deviation, thus a positive square root of , is sometimes called a standard error of the mean, and it enables us to examine the accuracy of an unbiased estimator.

Based on the information stated above, a sample average standard deviation can be defined as follows:

$$\sqrt{D(\bar{y})} = \frac{\sigma}{\sqrt{n}}. \qquad (2)$$

This characteristic cannot usually be defined precisely in practise as we do not know the population variance. Therefore, it is important to replace the unknown value by its point estimation , and we get:

$$odh\sqrt{D(\bar{y})} = \frac{s_y}{\sqrt{n}}. \qquad (3)$$

Since we were working with a population, and the characteristics of the population were available, we could calculate the standard error of the mean directly. *Table 1* shows selected characteristics of the population, which were calculated within the frame of the previous activities described in [1].

Tab. 1 Selected characteristics of the population (own calculations)

Population	μ	σ	$\sqrt{D(\bar{y})}$
	9.20197	**5.574498**	**0.244458**

That, apart from other things, allowed us to compare this measurement with its estimators obtained by particular samples. *Tables 2, 3, 4* show an overview of the obtained results and selected characteristics of the sample described in [1].

Tab. 2 Selected characteristics of the simple random sample (own calculations)

Sample	SRS – systematic			SRS – with help of random numbers		
	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$
Sample 1	9.15269	5.43582	0.238376	9.22019	5.70719	0.250277
Sample 2	9.06769	5.50411	0.241371	9.10981	5.54754	0.243276
Sample 3	9.13635	4.96938	0.217922	9.12923	5.19733	0.227918
Sample 4	9.14038	5.73322	0.251418	9.17962	5.73837	0.251644
Sample 5	9.78346	6.42324	0.281678	9.24788	5.51413	0.241810

Tab. 3 Selected characteristics of uniform and proportional allocation (own calculations)

Sample	Uniform allocation			Proportional allocation		
	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$
Sample 1	9.79962	5.75534	0.220511	8.88192	5.10039	0.192955
Sample 2	9.52788	5.40658	0.229614	9.53596	6.30391	0.235050
Sample 3	9.93827	6.02183	0.235961	9.22327	5.17975	0.198103
Sample 4	9.93871	5.98349	0.248117	9.12923	5.32163	0.202882
Sample 5	9.67692	5.64627	0.220972	8. 69404	4.90830	0.186032

Tab. 4 Selected characteristics of optimal allocation and two-stage cluster sampling (own calculations)

Sample	Optimal allocation			Two- stage cluster sampling		
	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$
Sample 1	9.65962	5.29517	0.186848	9.40365	5.85226	1.040275
Sample 2	9.49462	6.26057	0.206682	9.46577	5.74643	0.937844
Sample 3	9.46154	5.71500	0.204395	10.5004	6.18511	1.496538
Sample 4	9.24250	5.29480	0.198942	9.40192	5.70257	1.048770
Sample 5	9.73115	5.79131	0.206195	8.52788	4.65322	1.346991

As *Tables 2, 3,* and *4* show, the results obtained by the simple random sample with the help of random numbers come the closest to the actual standard error of the mean. The second closest result is from the systematic sampling, followed by uniform allocation, then proportional allocation, and then the optimal allocation. The two stage cluster sampling differed significantly.

Another criterion used to compare the taken samples was the mean deviation of each sample mean from the actual mean of the population. The mean deviation is calculated as follows:

$$\text{Mean deviation} = \sqrt{\frac{\sum_{i=1}^{k} (\bar{y}_i - \mu)^2}{k}}, \quad (4)$$

where $\bar{y}_i$ are individual sample means and k is a number of samples.

Table 5, which is also a result of the activities published in [1], shows a comparison of three kinds of random sampling.

Tab. 5 Comparison of different kinds of random sampling with the help of mean deviation. (own calculations)

Characteristic	SRS		Stratified sampling			two-stage cluster
	systematic	*random numbers*	*proportional*	*uniform*	*optimal*	
Mean deviation	0.2708	0.0578	0.3091	0.6011	0.3589	0.6768

The results presented in *Table 5* show that the simple random sample with the help of random numbers accounts for the best values. The average difference of sample means from the actual mean is only 0.0578, the second best is the systematic simple random sample with its value 0.2708. The two-stage cluster sampling demonstrated the biggest difference 0.6768. The uniform allocation of stratified sampling shows the second worst result, where the mean deviation is 0.6011.

Another criterion used to compare the given samples was their relative gains from stratification, which S.L. mentions in [page 77, 5]. It is weighing up the variance of stratified sampling and the variance of a simple random sample of the same size. More significant relative gains from the stratification was noticed only in two cases, namely when comparing proportional allocation of stratified sampling with the systematic simple random sample (0.9025) and the simple random sample with the help of random numbers (0.9434).

3 A new perspective on stratified sampling

After publishing the above results of our research, we were concerned with correct defining strata in stratified sampling. As stated hereinbefore, the quality of the results obtained based on the stratified sampling is contingent on the correct definition of strata. The regions of the Czech Republic were considered as strata in our last research, it means that we followed the formal organization. As the results demonstrate in [1], regions are not exactly ideal groups. The variability within them is relatively high and they do not differ a lot among themselves. Now, we focused on the fact how differently strata for stratified sampling can be defined, so that the obtained results would be more satisfying than in the previous case.

We took in consideration the fact that unemployment rate can be different in differently-sized municipalities, therefore we made a decision that a new grouping criterion will be the size of a municipality given by the number of inhabitants. We borrowed the categorization from the Czech Statistic Office – see e.g. [8], where we differentiate the following groups:

- municipalities up to 199 inhabitants;
- municipalities with 200 – 499 inhabitants;
- municipalities with 500 – 999 inhabitants;
- municipalities with 1,000 – 1,999 inhabitants;
- municipalities with 2,000 – 4,999 inhabitants;

- municipalities with 5,000 – 9,999 inhabitants;
- municipalities with 10,000 – 19,999 inhabitants;
- municipalities with 20,000 – 49,999 inhabitants;
- municipalities with 50,000 – 99,999 inhabitants;
- municipalities with 100,000 inhabitants and more.

Table 6 shows the number of municipalities in the Czech Republic, in each category.

Tab. 6 Number of municipalities in each size category according to the number of inhabitants (the CZSO and own calculations)

Category	**Up to199**	**200-499**	**500- 999**	**1,000 – 1,999**	**2,000 – 4,999**	**5,000 – 9,999**	**10,000 – 19,999**	**20,000 – 49,999**	**50,000 – 99,999**	**above 100,000**
Number of municipalities	1 608	2 012	1 304	678	376	138	69	42	16	5

We did not see uniform spreading of the stratified sampling meaningful due to the number of municipalities in each category. Therefore, we carried out only proportional and optimal allocation of stratified sampling – 5 samples from each kind.

Firstly, we created 5 proportional samples. The sample size in each stratum was defined according to (see page 17, [13]):

$$n_h = n\frac{N_h}{N}, \tag{5}$$

where

n_h is a sample size in h stratum,

n is a total sample size,

N_h is a h-stratum size,

N is a population size.

Table 7 shows the number of municipalities selected in each category, based on the number of inhabitants.

Tab. 7 Number of municipalities in the sampling in each size category according to the number of inhabitants at proportional allocation (the CZSO and own calculations)

Category	**Up to 199**	**200 -499**	**500 -999**	**1, 000 – 1, 999**	**2, 000 – 4, 999**	**5, 000 – 9, 999**	**10, 000 – 19, 999**	**20, 000 – 49, 999**	**50, 000 – 99, 999**	**above 100, 000**
Number of municipalities	134	167	109	56	31	12	6	4	1	0

As we can see, the last category, municipalities with more than 100,000 inhabitants, was not represented in the sample. We presume that this fact could cause less accurate results than if all the categories were included in the sample. *Table 8* shows the calculated selected characteristics from all 5 samples.

Tab. 8 Selected characteristics of proportional and optimal allocation (own calculations)

Sample	Proportional allocation			Optimal allocation		
	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$	$\bar{y}_i$	s_i	$odh\sqrt{D(\bar{y})}$
Sample 1	9.25510	4.98647	0.208174	9.17675	5.12996	0.205640
Sample 2	9.23478	5.30185	0.221910	9.10845	5.78815	0.230998
Sample 3	9.88580	5.88114	0.246896	9.64418	5.81105	0.244301
Sample 4	8.90485	5.55757	0.233371	9.16308	5.52874	0.225410
Sample 5	9.41249	5.66548	0.236503	9.38354	5.86010	0.232929

Standard error of estimated mean in stratum *h* is calculated according to (presented in e. g. [3]):

$$odh\sqrt{D(\bar{y})} = \sqrt{\frac{1}{N^2}\sum_{h}^{L}\left(\frac{N_h^2}{n_h} - N_h\right)\cdot s_h^2}, \tag{6}$$

where

N is a population size,

N_h is the population total in stratum h,

n_h is the sample size in stratum h,

s_h^2 is an estimator of the population variance in stratum h (for sampling with and without replacement).

The estimator of the population variance in stratum *h* s_h^2 is defined as:

$$s_h^2 = \frac{\sum_{k}^{n_h}(y_{hk} - \bar{y}_h)^2}{n_h}, \tag{7}$$

where y_{hk} is a value of *k*-th unit in stratum *h* and $\bar{y}_h$ is an estimator of the population mean in stratum *h*.

After proportional allocation of stratified sampling, we carried out optimal allocation. The optimal sample size in stratum *h* (for sampling without replacement) is determined by the relation presented in e. g. [3]:

$$n_h = n \cdot \frac{N_h \cdot S_h}{\sum N_h \cdot S_h}, \tag{8}$$

where

N_h is the population total in stratum *h*,

n_h is the sample size in stratum *h*,

S_h is the population standard deviation in stratum *h*.

The standard deviation S_h is calculated according to the formula:

$$S_h = \sqrt{\frac{\sum_{k}^{N_h} \left(Y_{hk} - \bar{Y}_h\right)^2}{N_h - 1}}, \tag{9}$$

where

Y_{hk} is a value of *k*-th unit in the population,

N_h is the population total in stratum *h*,

$\bar{Y}_h$ is the population mean in stratum *h*.

We calculated selected characteristics on a base of five random samples – they are also presented in *Table 8*.

Focusing on the comparison of these samples, we have to supplement comments with the calculations of the average deviation of sample means from the population mean. *Table 9* contains our calculations.

Tab. 9 Comparison of proportional and optimal allocation of stratified sample with the help of the average deviation (own calculation)

Characteristic	**Stratified sampling**	
	Proportional allocation	**Optimal allocation**
Average deviation	0.3476	0.2190

The results in the previous tables show that standard errors are smaller in the case of optimal allocation of stratified sampling but the differences are not so significant. If we compare the new samples with the previous ones (presented in [1]), it is evident that new allocation of strata does not bring any benefit because the standard errors are greater than in the case when the strata were defined as regions of the Czech Republic.

Let us look at the values of average deviations. The average deviation is considerably smaller in the case of optimal allocation of stratified sampling in comparison to proportional allocation. We can also notice by the comparison of the new results with the results from the previous research (see *Table 5*) that the average deviation for optimal allocation is significantly smaller in comparison to the stratified sampling when the strata were defined as regions of the Czech Republic. We can even register that it is the second best result (the best is SRS using the random number generator). The value of this characteristic is now worse for proportional allocation of stratified sampling than in the previous research.

We omit the comparison with the help of the relative gain from stratification because it does not bring significant benefits which would make the decision on a kind of sampling easier.

Conclusion

The calculations and comparisons of the various kinds of sampling mentioned above indicate that neither stratified sampling nor two-stage cluster sampling improve the quality of estimates. The average deviation shows that the estimations obtained by SRS do not differ from the population characteristics as much as the estimates obtained from other kinds of samplings. We achieved a certain improvement in the quality of estimations, with respect to this criterion of the comparison, by changing the definition of strata from the "regions of the Czech Republic" to the "municipal size categories". However, we did not get as a significant improvement of the estimations quality by changing the strata definition as we had expected.

The reason can be unemployment rate blindness in relation to the number of inhabitants, or larger variability of values within the strata, and small variability among them. So, we assume that municipal size categories are not a suitable sorting criterion either.

In conclusion, let us add that it is necessary to take into consideration the fact that we are limited by many various factors when carrying out sampling in practise. Firstly, it is the availability of data, which are not often in such a structure that can be subdivided into suitable subgroups. Furthermore, there is the means, which makes us minimize the survey costs. It is necessary to harmonize all these requirements and choose a suitable compromise. Even though it is evident that the best way of taking random samples from a population would be simple random sampling in relation to the estimations quality, its financial and organizational demandingness makes us use some of the more complex kinds of sampling. Two-stage cluster sampling is very often the most common solution in practise. It allows us to reduce survey costs but the efficiency of estimators is small. We have here presented the characteristics and results of two-stage cluster sampling with equal probabilities only. We could have obtained more efficient estimators if we had carried out sampling with unequal probabilities. This idea can be an impulse for the next research.

Literature

[1] GURINOVÁ, K.; HOVORKOVÁ VALENTOVÁ, V. Možnosti provedení náhodných výběrů z populace ČR za účelem zkoumání vývoje hospodářských ukazatelů. *In VII. ročník mezinárodní konference aplikované statistiky FernStat_CZ 2010.* Ústí nad Labem 23. – 24. 9. 2010. Sborník příspěvků. Ústí nad Labem: UJEP, FSE, 2010 – not printed yet.

[2] ČERMÁK, V.; VRABEC, M. *Teorie výběrových šetření*. Část 1. 1. vydání. Vysoká škola ekonomická v Praze, Praha, 1999. ISBN 80-7079-191-8.

[3] ČERMÁK, V.; VRABEC, M. *Teorie výběrových šetření*. Část 3. 1. vydání. Vysoká škola ekonomická v Praze, Praha 1999. ISBN 80-245-0003-5.

[4] KAHOUNOVÁ, J. *Praktikum k výuce matematické statistiky I. Odhady*. 1. vydání. Vysoká škola ekonomická v Praze, Praha, 2000. ISBN 80-245-0070-1.

[5] LOHR, L. S. *Sampling: Desing and Analysis*. 2.vydání. Brooks/Cole, Boston (USA) 2010. ISBN 978-0-495-11084-2.

[6] PACÁKOVÁ, V. a kol. *Štatistické metódy pre ekonómov*. Prvé vydanie. Iura Edition, Bratislava, 2009. ISBN 978-80-8078-284-9.

[7] PECÁKOVÁ, I; NOVÁK, I.; HERZMANN, J. *Pořizování a vyhodnocování dat ve výzkumech veřejného mínění*. Vysoká škola ekonomická v Praze, Praha, 2000. ISBN 80-7079-357-0.

[8] Český statistický úřad: *Tab. 2.3.1 Průměrná obytná plocha /v m2/ na 1 byt podle počtu obytných místností a podle velikostní kategorie obce k 1. 3. 2001 [online]*. Praha, Český statistický úřad, [cit. 2010-10-14]. Available from WWW: <www.czso.cz/csu/2005edicniplan.nsf/p/4131-05>

Ing. Kateřina Gurinová, Ph.D.
Ing. Vladimíra Hovorková Valentová, Ph.D.

ZPŮSOBY PROVÁDĚNÍ NÁHODNÝCH VÝBĚRŮ ZE ZÁKLADNÍHO SOUBORU PRO POTŘEBY ANALÝZY EKONOMICKÝCH UKAZATELŮ

Cílem příspěvku je poskytnout přehled možností, jak provádět náhodné výběry z populace České republiky, a jejich následné porovnání. Porovnání předložených metod bylo provedeno pomocí vybraných statistik, které umožňují zvolit metodu, jež přináší nejpřesnější odhady. Při závěrečných doporučeních však nesmíme zapomínat také na skutečnost, že při provádění náhodných výběrů v praxi jsme limitováni i finančními prostředky, pracovními silami, zabývajícími se šetřením, a dalšími faktory. Předkládáme tak v prvním případě řešení ideální a ve druhém optimální, a to s přihlédnutím ke všem ovlivňujícím faktorům.

DER VORGANG DES ZUFÄLLIGEN HERAUSLÖSENS AUS DER GRUNDGRUPPE FÜR DIE ANALYSE DER ÖKONOMISCHEN PARAMETER

Das Ziel unseres Artikels besteht in der Absicht, Ihnen eine Übersicht verschiedener Methoden, getroffen durch eine zufällige Auswahl aus der Bevölkerung der Tschechischen Republik, zu zeigen sowie diese Methoden miteinander zu vergleichen. Der Vergleich dieser Methoden wurde anhand ausgewählter Statistiken durchgeführt. Diese Statistiken ermöglichen es, diejenige Methode auszuwählen, die die genaueste Schätzung bringt. Bei unseren Empfehlungen darf man die Tatsache nicht vergessen, dass wir bei der Durchführung des zufälligen Herauslösens in der Praxis durch finanzielle Mittel, Arbeitskräfte, und weitere Faktoren limitiert sind. Somit können wir im ersten Fall die ideale und im zweiten Fall die optimale Lösung aufzeigen, die alle beeinflussenden Faktoren berücksichtigt.

SPOSOBY PRZEPROWADZANIA PRÓB WYRYWKOWYCH Z ZESTAWU BAZOWEGO NA POTRZEBY ANALIZY WSKAŹNIKÓW EKONOMICZNYCH

Celem artykułu jest przedstawienie możliwości przeprowadzania prób wyrywkowych z populacji Republiki Czeskiej oraz ich porównanie. Porównanie przedstawionych metod zostało przeprowadzone za pomocą wybranych cech, umożliwiających wybór metody dającej najdokładniejsze szacunki. Przy końcowych zaleceniach nie można jednak zapomnieć, że podczas przeprowadzania prób wyrywkowych w praktyce istnieją ograniczenia związane zarówno ze środkami finansowymi, jak i siłą roboczą zajmującą się badaniem oraz innymi czynnikami. W artykule przedstawiono propozycję idealnego rozwiązania oraz rozwiązania optymalnego, przy uwzględnieniu wszystkich czynników na nie wpływających.