Ranking of Sampling designs

Roland Szilágyi
Ph.D associate professor
University of Miskolc, Faculty of Economics

This study deals with the quality and with some questions of the qualification of datas of samples formed on the basis of different sampling methods, sampling plans, and of estimations based on these datas.
Thereinafter I have undertaken for the elaboration of principles, on the basis of which – if not in the point of all interpretable criterias of quality, but on the basis of some important parameters - datas and estimation results gained from different samples can be qualified, ranked. Of course, I have made an attempt to this beside the existence of several theoretical conditions.

THEORETICAL CONDITIONS OF THE STUDY

First of all, I assumed the total realization of the sample, that all the asked relevants give answer to all the asked questions. This condition makes it obvious that I did not tried for the estimation of the measure of total error. Beside the 100 percent rate of response I assumed that other errors committed by the interrogator and the errors committed in the course of data recording, together with nothing more non-sampling error do not bias the result.
I did not take into consideration the intention of sampling plans to ensure cost-efficiency in any sense, neither during the planning, nor the interrogation, nor the registration, nor the evaluation of datas.
As we deal with a kind of special experimental study, in this case the selection and the realization of the sample does not have costs, so the unit costs neither can be defined, nor can be present in the study.
Ensuring uniform pressure on the investigators working in certain sessions, nor any other –neither economic, nor time saving – criteria is not my goal. In a spirit of these conditions I tried for the study of the representative efficiency of sampling in the aspect of estimation results.

The base of the experiments, the baseline hypothesis

During the work, I generated samples from the population on the basis of 53 different sampling procedure – form the database of Household Budget Recording (Háztartási Költségvetési Felvétel) of 2005 – in order to analyze in details the effect of the sampling plans on the result. I took samples in three different sizes: samples of 30, 150 and 900 households were chosen. It provided an opportunity for the control of my inferences and estimation results – in practice does not fulfilled condition – that I possessed the information of the population of the examined phenomenon. Samples were made by simple random, stratified, and stratified by more criteria sampling methods, because according to my experiences in business studies difficult sampling

DOI: 10.26649/musci.2015.095

methods are not used. The variables involved in the stratification got out from the followings:

- Age,
- Income,
- Useful floorspace of the house,
- Number of rooms,
- Number of cars,
- Number of televisions,
- Region,
- Consumption,
- Activity status of the head of the family,
- Size of the family,
- Gender,
- Type of household,
- Population density,
- Population of the settlement.

In the course of the examinations I calculated the estimative functions and their standard deviation for all the generated samples. The point estimation – in this case the mean – standard deviation, otherwise the standard error of the estimation which has a well-known significance role in the judgement of the estimation's accuracy and reliability, as well as it is essential to make interval estimations. However, just in itself is unsuitable to help with compare more independent samples so that to decide which sample is more suitable for the estimation of the examined parameter. Therefore for the further analysis I had to study the characteristics of the sample in view of rank and classify.
Despite the fact that independent samples were used, but the examined variables were the same (Total Consumption of Households) in every cases and the population was the same, according to the variance of the population, - which in practise seem to be normally stable – we also can not have an effect on the results of the sampling plan's comparison.
It's obvious to study the most flexible characterictic of the samples which is the size of the sample, it means according to the description above realize in various sizes. The given results of the different sampling methods will certainly revel to the well-known hypothesis which sound the advantageous properties of the larger samples.
Considering that several sampling procedure, sampling plan were used, I definetely had to mention the effect on the examination regarding to the sampling plan of estimation. The quantify of the sampling design's effect is exceedingly written by [1], [3], [2], and several other studies. Measuring the effect of the sampling plan, based on the dissolve of the standard deviation and the variance of the estimative function, and it can be determined by using the following formula:

$$DEFF = \frac{SE_A^2}{SE_{EV}^2}$$

Where in the numerator we can find the variance of the calculated parameter according to the actual sampling method, the denominator has the same size, but it is calculated in the case of the simple random sampling variance. This indicator based

on practically the classic variance's decomposition, because it is well-known that the square of the mean estimation of the standard error – for exampe in the case of stratified sampling – is equal to the internal standard deviation's square, while in the case of the selection of the appropriate estimation of the simple random sampling the variance of the layer group added to the individual variances.
The evaluation of the indicator is considered to be easy, because the values which are bigger than one shows that the estimation based on the given sampling plan is less effective (has higher standard deviation) than the estimation made by simple random sampling plan.Of course the result which is less than one indicates more effective estimation. [1]
On the basis of these facts can be suppose to be that in the view of both the efficiency and the accuracy, those samples which made by more detailed information ensure more relevant results, in these samples where for the ply more and examined changing relationship of stochastic variables are included in the sampling plan.

APPLIED SAMPLE PARAMETERS

I ranked the examined samples with regards to the following parameters:

- effect of the sampling plan/Design effect – Deff
- Coefficient of Variation – CV, incase of mean:
- effective samplesize

By effective samplesize, I mean the following:
knowing the value of Deff, it can be shown that how bigsample we should choose with regard to the originial sample toget the same estimated results. So, with a Deff value bigger than 1, the effective sample size indicates that using a better sampling plan how much bigger or how much smaller sample we should take in order to get the same results.
I experienced that the effective sample size can be considered as a guideline when ranking the sampling plans, which ensures the ranking in regard to the other 2 viewpoints.

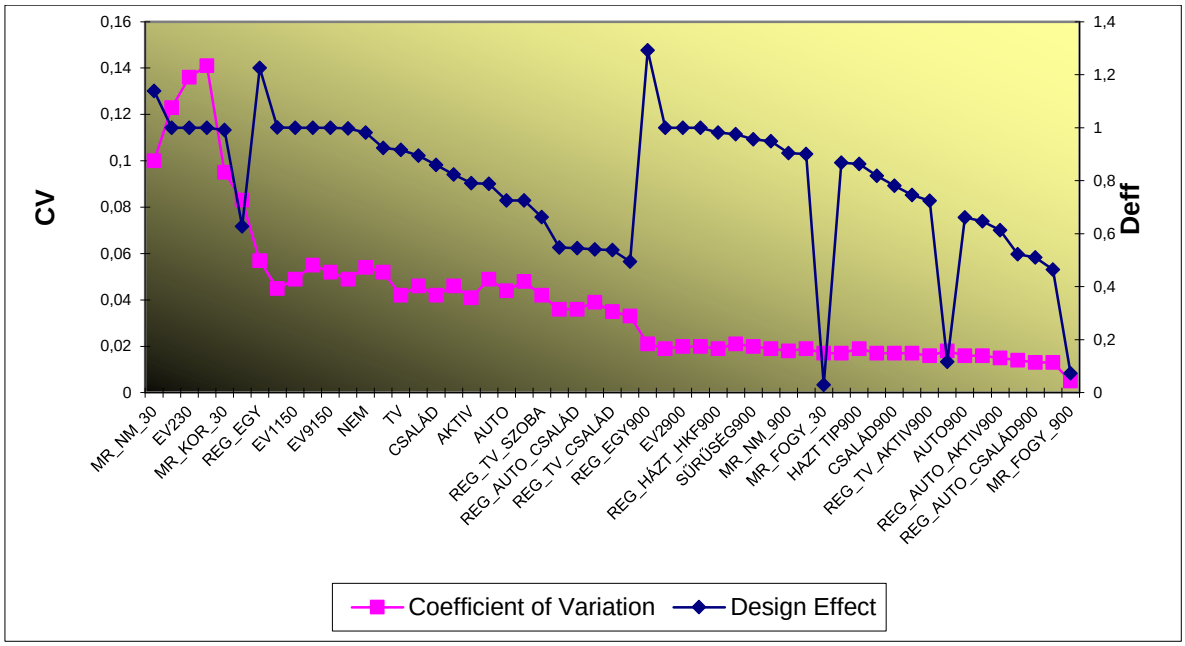


Figure 1.: the parameters of the households' estimated whole consumption

The illustration well indicates that according to the sample sizes, the samples form 3 different groups. There are 3 artificially stratified sample exceptions, which have been stratified based on the realvalues of consumption., according to the deciles. At these samples ,the value of Deff is extremely low and very much closeto 0. The artificially made samples – created according to the income variant which is highly correlant to the consumption – have a lower value than the other samples' Deff.
When examining the relative standard error, we can state that the smaller samples have worse results than the bigger ones. The effect of the sampling plan shows better tendency with more complex samples in case of every samplesize.
Layering according to more parameter simplies better results, supposing that the layers are not to ofrittered, because in that case, we are heading towards a result of a simple random selection with unnecessary work. The illustration shows that those layering parameters are good which include more scalefactors which are in stochastic relation with the examined parameter (inourcase, the consumption) – such as the number of cars, the number of rooms or the size of the household.

THE EXPERIMENTAL RESULTS

The literature on sampling theory contains rather little information on the Deff index. It obviously indicates the less effective sampling plans but there is only indirect information on the sampling plans that ensure lower variance. Based on all these, we can presume that the Deff index indicating the effect of the sampling plan on the estimated results, measures the effect of the sampling plans that are less efficient in regard to the simple random sample, but regarding the same size and same variant, we do not have to decide in the case of estimate function, which is the more efficient.

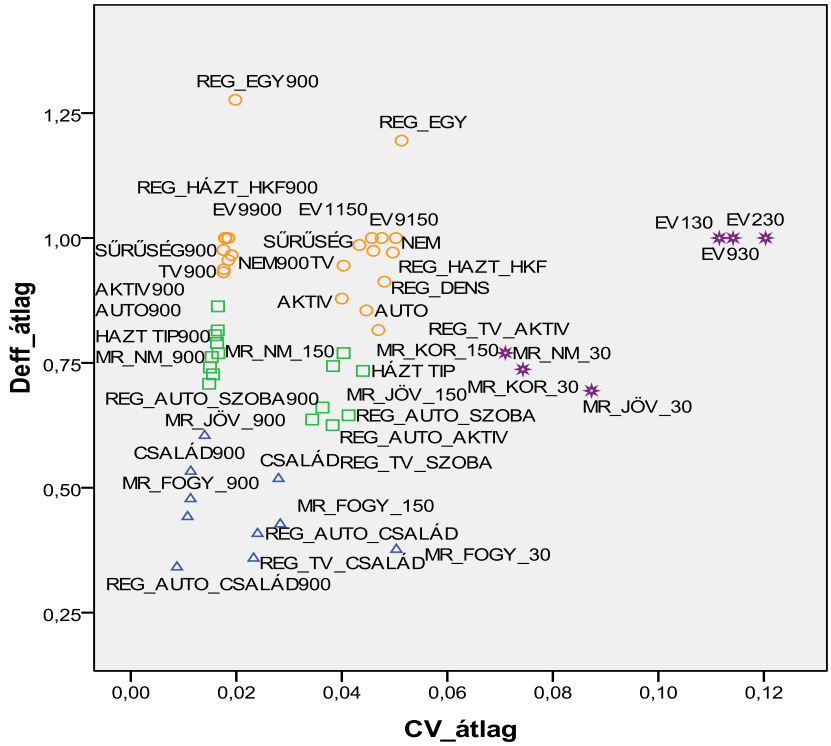


Figure 2: Results of the cluster analysis carried out on samples

Fort he classification was used cluster analysis based ond Ward's methede and Euclidien distances. [4] I have to say that the figures transparency i worse because of the annoying crowdness, but I feel it much needed to show the names of the figures so it's easier to identify them. From the figure its easy to see that the low scale of EV samples are bad based on both variable, since it has extraordinary large Deff value and high CV value aswell. In the same figure (shown as purple star), less effective region also contains small sized artificially stratified samples where the samples size probebly limited the layerings pozitive effects.

The best group is presented with a blue triangle, multiply stratified, and it contains bigger scales of artifical samples.

We can see that the top of the figre there is the least effective uniform stratification (REG_EGY) containing samples where the Deff value is bigger then 1, these where succesfully revealed by clastering.

The literature says that it's acceptable for a Deff indicator to show what sampling is wosre then another at a simple random sampling but in case of a better sampling like a stratified one it give questionable results.

There is no doubt that the professional literature, and the statistics used in the ratio Deff practical use, evaluation is mostly based on the estimated sample group occurs. However, it is recognized that in today's computing infrastructure group of samples have very little beneficial properties. According to some of the literature its specifically a benefit - and here I must point out that the practical aspects of the nation-wide sample to sample extremely useful - feature known as the candidate of the data collection phase of cost savings. Since the present study does not present anything else then the cost over sampling, we can leave out this positive option. The results of cluster analysis, however, reveals that the ratio Deff plays a significant role

in the settlement samples of groups of efficiency. Despite the fact that only a few samples exceed the value unity when the interpretation is commonly known and used determined . According to Figure 2, however, it has to be set on the vertical axis there is much stronger separation of sample groups, such as along the horizontal axis, according to the CV. Examined how the two trained groups are formed on the basis of the characteristics of the typical average.

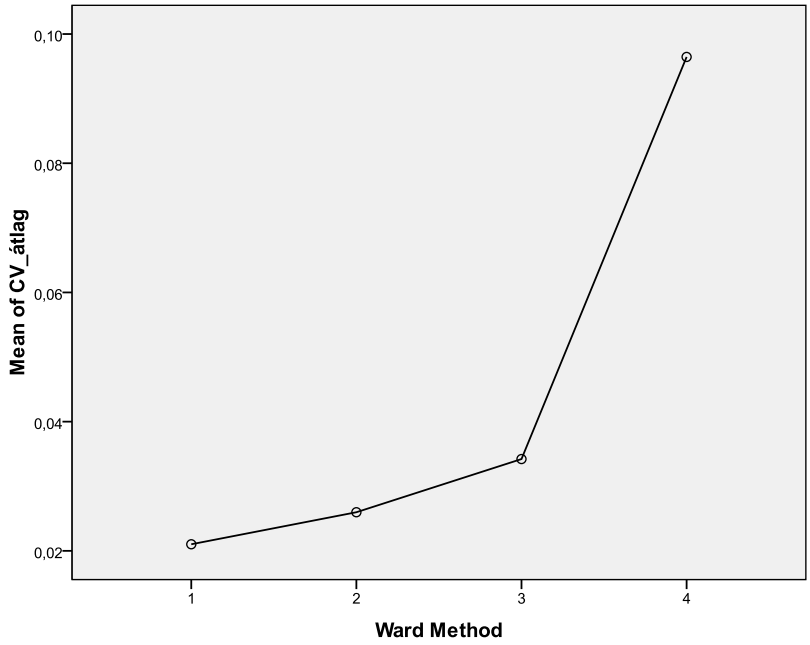


Figure 3a: Evolution of the average CV for each sample group

Based on the 3a figure, the relative standard error is the average of an exponential picture, ie the samples belonging to the fourth group is very high relative error of the estimated parameter. For the first three groups, however no significant differences in these samples, i.e., the relative error in itself does not provide adequate guidance for preferential status.

The 3.b. figure, which shows the ratio Deff group averages, there are significant differences between the groups. For each number of clusters as was the average value Deff listed below 0.5, which is twice as good matter basis classical interpretation of the pointer as a simple random sample of the same size.

No clear conclusions can be drawn on the basis of averages of groups 3 and 4, as shown in the figure, the fourth sample group average Deff least effective is kept lower than the third average. However, this minimal distractions we navigate the rigorous conditions of the cluster analysis. Both said they are in a cluster - can be considered as outliers - values that the cluster centroids can be pushed along the vertical axis. These small sized simple random (EV) samples, the estimation results are limited reliable and uniform stratification samples, referred to already.

However, it turns out from the results of the cluster analysis, that the Deff indicator bears a significant part in grouping the samples by their efficiency. Despite the fact, that it exceeds the unit value only in some cases, when the interpretation is well-known and its application is determined. By the Figure 2. I must claim, that along the vertical axle the sample groups are more starkly differentiated as along the horizontal

axle, namely by the CV. I analysed, that how the features of the mean form in groups formed by two features.

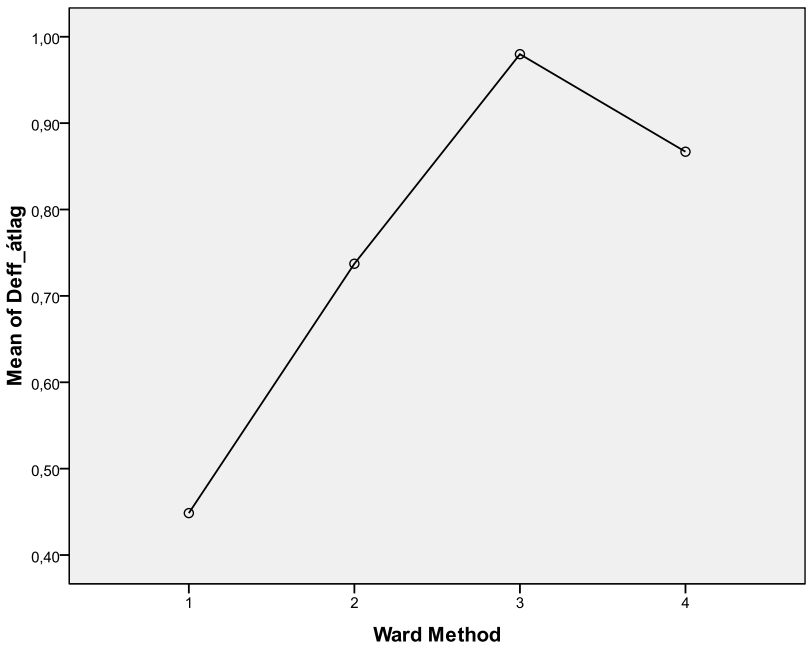
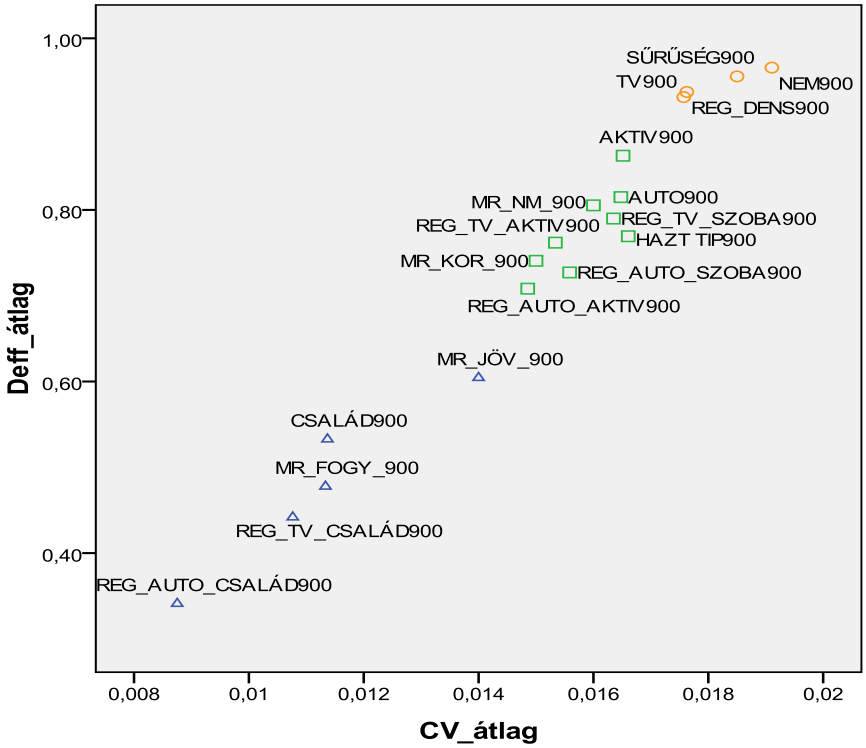


Figure 3.b.:Trends of the means of Deff in each sample groups

In the previous subsection I highlighted the assumption, which is that from the bigger samples we can get better estimation results. I maintain this theory, but I study, that what kind of conclusions can we solely draw from concentrating on the bigger samples. As a result of this, I only displayed the 900 element samples in the average Deff and average CV dimension. Maintaining the fact, that the cluster analysis has been successful, so the samples can be ranked among different homogeneous groups (measured by the help of the Deff and CV indicators) by their efficiency. In other words, getting rid of the extremely small sample size and the effects of samples resulted by the less useful stratifying procedure, the outlines of the clusters are outlined differently.



4.a. graph:The samples' clusters with 900 elements

In the graph 4.a we can see, that if the extreme fluctuation, caused by the different samples ceases, the samples will be placed on the diagonal which is between the 2 dimensions. It means that both of the dimensions play a proportionately important role in shaping the clusters. The Deff and CV are good coupling because they are in connection with each other caused by the relation between the relative error and the converse quadratic number of elements. The Deff gives us new results which are independent from the samples.

According to the specialized books the mathematical and statistical method demonstrate the features of the Deff indicator's ability to show how many times the sampling plan better or worse is than a simple random sample at the same size. During my empirical researches I also revealed that the Deff indicator together with other indicators can make an efficiency ranking of sampling plans which are better than a simple random sample.

Acknowledgement: This research was carried out in the framework of the Center of Excellence of Sustainable Resource Management at the University of Miskolc.

BIBLIOGRAPHY

[1] KISH, L: **Kutatások statisztikai tervezése**; Budapest, 1989.

[2] MARTON Á.: **A reprezentatív felvételek megbízhatósága** 1991.

[3] MARTON Á. – MIHÁLYFFY L.: **A mintavételi hiba kiszámításának néhány kérdése**; Statisztikai Szemle, 1988. 4. sz. pp. 350-366.

[4] VARGA B. – SZILÁGYI R.: **Quantitative Information Forming Methods**; TAMOP – 4.1.2-08/1/A-2009-0049 Virtuális vállalatok; 2011. http://elearning.infotec.hu/ilias.php?baseClass=ilSAHSPresentationGUI&ref_id=2774