UNIVERSITY OF MISKOLC

FACULTY OF ECONOMICS

ROLAND SZILÁGYI

TREATMENT OF SAMPLE-BASED ESTIMATE ERROR: PROBLEMS ARISING FROM NON-RESPONSE

PH.D. THESIS

MISKOLC

2011

UNIVERSITY OF MISKOLC

FACULTY OF ECONOMICS

ROLAND SZILÁGYI

TREATMENT OF SAMPLE-BASED ESTIMATE ERROR: PROBLEMS ARISING FROM NON-RESPONSE

PH.D. THESIS

NAME OF DOCTORAL SCHOOL:	**VÁLLALKOZÁSELMÉLET- ÉS GYAKORLAT DOKTORI ISKOLA**
HEAD OF DOCTORAL SCHOOL:	**DR. SZINTAY, ISTVÁN** *professor* *CSc*
ACADEMIC TUTOR:	**DR. BESENYEI, LAJOS** *professor* *CSc*

MISKOLC

2011

CONTENTS

1. THE PURPOSE OF THE RESEARCH, THE STRUCTURE OF THE DISSERTATION

Due to alterations in market structure, the rapid changes in the environment and the development of information technology, the time necessary for preparing a decision has decreased significantly. The decision-makers of business life can only be effective in competition with both time and competitors if they constantly improve the techniques used during the preparation of decisions. There is a wide choice of quantitative analytic methods to make possible the inspection of the characteristic features of the individuals. However, choosing and applying the adequate quantitative method is not the only key to success. The other important factor must be the reliability of the data used to obtain information.

Research based on samples and their conclusions play an increasing role in making business decisions and also in creating information. The spread of sampling is mainly due to the lower expenses and shorter time needed for an investigation. Surveys based on samples are becoming more popular not only on a micro-level, but also in case of macroeconomic investigations. However, the spread of sampling has a great risk because of the quality of the samples. Information gained from sample data can include a variety of mistakes. The exploration and evaluation of these mistakes with mathematical-statistical methods, together with the reduction of the negative effects they have on the results, constitute the purpose of my research. My aim is to work out an analytic system, a system of points of view, which adapts to the opportunities provided by information-gathering techniques and technologies and which complement the currently used methodology.

As the priority of my work I chose the exploration of the possible faults of surveys based on samples and the negative effect they have on the result. Consecutively,I searched for solution variants for handling of mistakes, primarily for non-random errors. The most effective way of handling mistakes is to prevent them from happening. If the error has already occurred, then as the first phase of the treatment we have to map the reasons for the mistake. To what reasons can the mistakes of social science surveys be traced back? What could be the deriving factors which might lead to the categorization of a conclusion as being unfounded? The specialized literature gives the following answers to these questions: unreliable data, unfounded conclusions, based on too small a sample, the relative errors are high, the insignificance of the effect on the original issue, etc.

Further critical factors might come up in a case when the research processes, analyses and conclusions are drawn from a primer information set. The possibility of making a mistake increases during sampling. The mistakes committed during

sampling can result in bias in the quality and quantity of the produced source data, which is dangerous, because false basic data can lead to improper conclusions even with the best methodology and the help of professional analytic techniques and methods. I searched for answers and solutions to the aforementioned problems and questions by analyzing the database of Household Budget Survey (HBS) submitted by the Hungarian Central Statistical Office. To prove the authenticity of the data I presented the resources of HBS information, and the methodology of the intake. I aspired to introduce the effects of errors and a scientific explanation for the decrease in negative effects through the analysis of different samples.

In the first part of my dissertation I examine the circumstances of forming the sample on which the conclusions are based. In the first step I introduce the theoretical principles concerning sampling procedures and then I review the Hungarian and international scientific literature on sampling and error calculation methods, which helped in utilizing the research results.

In the further sections of my scientific work I deal with the identification and systematization of potential sources of errors, for which other disciplines besides a statistical-mathematical approach – namely, marketing and sociology – provided additional data. The experience of questionnaire surveys and opinion polls helped in recognizing the steps made in order to handle errors and the methods and conceptions used to decrease the error margin.

During the process I generated samples for the population from the HBS database - using 53 different sampling methods in order to study in more detail the effect of sampling plans on the results. This was to check my conclusions and estimates, as I had the population of information (which is not given in real practice). Samples are taken by simple randomised, stratified and complex stratified methods, as according to my experience more complicated methods are not used in corporate research.

Furthermore, I aimed at working out points of view based on which results and estimates gained from several samples can be ranked. Of course I tried this in the case of several hypothetical conditions.

Based on domestic and international research findings, we can state that the biggest problem when carrying out surveys is when answers are not given. Nowadays questionnaires with answer rates lower than 50% are not rare. The scale of the problem can be seen in the findings of Ay and Vita (1998) regarding experimental income conditions, according to which the preliminary request for participation faced approximately 90% refusal rates.

Obviously, selective answering not only reduces sample quantity but also increases the variants of estimates and the scale of bias. That is why in, the following sections

of my dissertation, I dealt with investigating one of the most important non-sampling error types: non-response errors. In this area I particularly focused on partial, item-level non-response. Using different analytical methods I analysed how large the impact of supplying these answers is on the results of descriptive models. Addressing this issue I generated non-responses of different degrees, considering the conception proved by several domestic studies (Keszthelyiné 2006, Havasi 1997, Havasi and Schnell 1996), that questions about income and consumption are only reluctantly answered by people with higher incomes, who make up a significant portion of non-respondents. Then I investigated whether the impact of a sample-taking plan can be recognized in the bias caused by non-response.

The last phase of my research was testing the role of model-based approaches in handling non-response. Using probability as a weighting factor I managed to reduce this bias. However, we should not forget that the aim of practical research is not only to mitigate the negative effect of bias, but also to obtain a more reliable estimate of population parameters.

Identifying tendencies plays a significant role in eliminating the bias caused by non-response. It is worth examining the differences between response and non-response criterion tendencies. In order to achieve success the sample should be grouped based on different criteria. Using these methods I created a model for estimating weighing tendencies, which limits the degree of bias to within an acceptable range.

2. THE METHOD OF RESEARCH

In this dissertation I deal not only with the determination of the potential faults of surveys based on sampling and the simple introduction of each type of fault, but I also work out processes that enable the minimization of mistakes and the reduction of bias. I do it in a generalized way, so that these processes could be of help to both individual researchers in social sciences and to research organizations.

For the examination of analyses based on data from samples, huge databases are the most desirable. Lacking material resources, I did the analysis based on reliable secondary data. However, I have to deal with the different methods of collecting data and observation, as they can generate different types of errors.

Throughout the analysis of data I carry out research on the database of the HBS (Household Budget Survey). In the course of this I intend to sort the extant results based on a new logical structure. I would like to use methods that can be utilized not only on a macro-level, but even on the level of small-size enterprises. Such enterprises are not in the possession of error calculating software and algorithms, unlike the representatives of official statistics. In absence of adequate methods they are not able to produce quality information. The methodological results of my empirical research aim at filling this gap.

During data processing, in order to carry out complicated and time-consuming methods, to perform mathematical calculations, and to present graphs and tables, I use the Windows-based SPSS statistical software version 17.0 and Microsoft Excel during processing.

3. THE NEW AND NOVEL STATEMENTS OF THE RESEARCH

Representativeness is a defined expectation in the planning of sampling and in choosing a sampling process, and not only in the case of stratified samples. In practice, representativeness is expected from the intake based on numerous parameters.

The drawback of representativenss based on several parameters is that based on the numerous criteria several strata and cross-classes are generated, the size of which can become unduly small. I examined the samples based on the following parameters in order to decide whether it is worth validating several parameters by composing representativeness expectations:

- Design Effect – Deff
- Coefficient of Variation – CV (relative variance of parameter),

  $$\frac{SE\hat{\Theta}}{\hat{\Theta}}$$
- Effective sample size, n/Deff

I defined the above-mentioned parameters along several variables to estimate the anticipatory value, examining whether a kind of rank prevails among the chosen samples.

Using the results obtained from estimating the percentage of household consumption per capita, the examined parameters of the samples are shown in Figure 1. Based on the figure, it might be hypothesized that the samples composed of more detailed data provide more relevant information regarding both effectiveness and accuracy. The stratification of samples utilizes several variables in a stochastic connection with the examined variable drawn in the sampling design. After the validation of the results with significance investigation and classification methods (cluster analysis), the following proposition can be stated:

Proposition 1

It is known from the specialized literature that valid results, those producing and publishing relatively few errors resulting from surveys based on sampling, can only be guaranteed if there are adequate databases. During my research I found that in Hungary the data necessary for correcting the results of population-based surveys based on sampling is widely dispersed and can be found (though not always accessed) in the data collections of different

offices and organizations. In order to improve the usefulness of research surveys, these data sources need to be integrated.

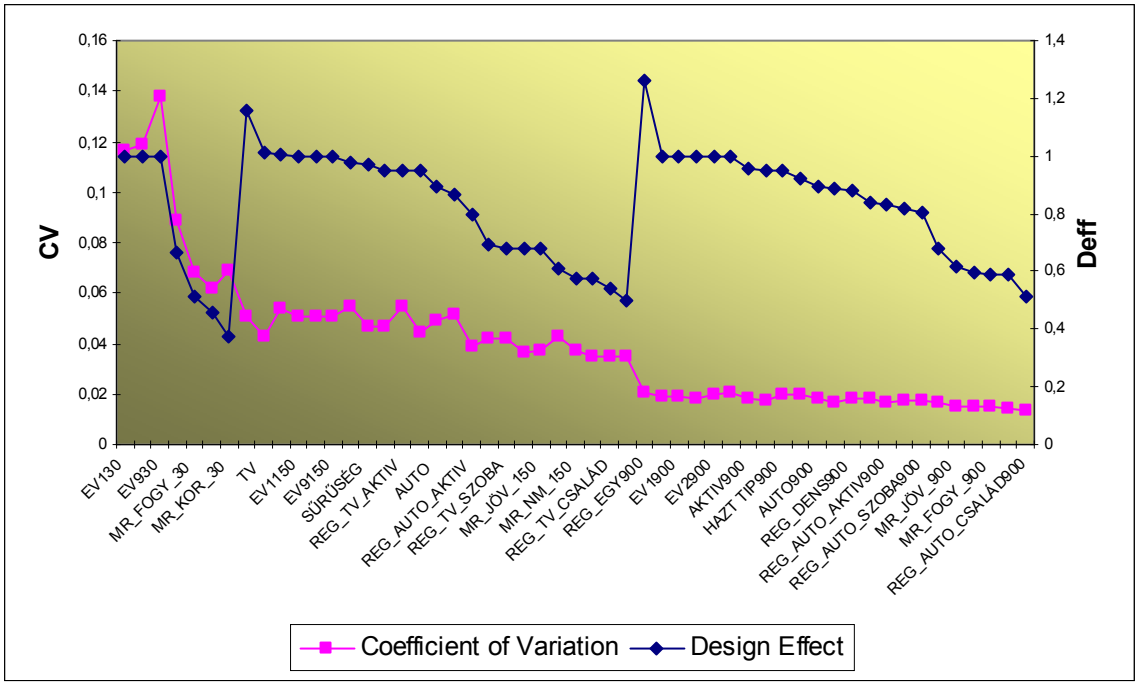

Figure 1. Estimation features of average household consumption per capita ranked by effective sample size

It supports the pertinence of proposition 1 that in surveys based on sampling it is not uncommon to use complementary information, which are mainly used to determine the sampling error, to increase the response ratio, and to reveal bias. See the studies of Estevao and Särndal (2002) and of Roy and Safiquzzaman (2006) regarding the use of two-phase samples.

Särndal and Lundström (2008) assign an indispensable role to complementary information in dealing with bias caused by non-response, emphasizing that the quality of the complementary information we are using can have a considerable impact. Complementary information also helps in methods used for estimating subclasses and groups, according to Estevao and Särndal (2004).

Figure 1 sheds light on the well-known assumption stressing the advantages of bigger samples. It is clear that the effective sample size can be seen as a guiding principle when ranking samples, which enables a ranking based on the other two aspects.

In order to improve the comparability, interpretation and demonstrativity of the results of ranking, I continued the ranking by leaving the size of the sample out of

consideration in order to prove that the aforementioned aspects are capable of setting up a rank among different sampling plans in the case of the same sample size. During the classification I did a hierarchical cluster analysis. The analysis dimension was given by the average of the Deff indicators of the different variables together with the average of the CV indicators, in which the 900- item samples make up the following groups.

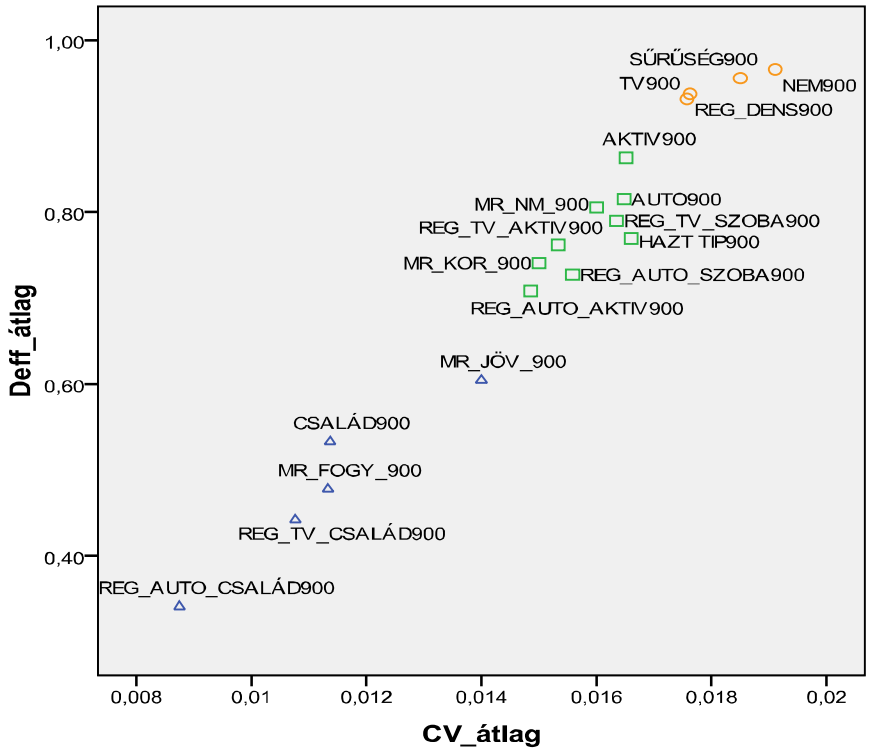

Figure 2. Clusters of 900-item samples

It can be seen in Figure 2 that if extreme fluctuation among samples of different sizes is eliminated, samples are situated along the diagonal. This means that both dimensions play an important role in forming clusters. Applying Deff and CV together seems to be a good match, because relative error and sample size are in close connection wih each other.

Proposition 2

We know from the literature that mathematical and statistical methods demonstrate the feature of the Deff indicator which compares a given sampling plan and and a random intake of the same size. In addition, my empirical research also reveals that the Deff indicator together with other indicators is able to yield an efficiency rank of sampling plans.

Non-response is one of the biggest problems occurring when surveys are taken. To investigate the effect of non-response I made estimates of the value of average entire consumption expenses based on different sampling processes and sampling plans. From the numerous samples the stratified sample MR_Fogy-900 proved to be the most effective. In this sample nearly 10% of the entire population sampled was chosen in strata proportional to consumption expenses deciles. Priority was given to efficiency and precision, so selection was based on the values of Deff and CV indicators. Efficiency of the sampling plan was 0.27 while the relative standard error was 0.5%.

In my experiments I examined different scales of non-response. It was undoubtedly proved that the increase of non-response impairs estimate values. Here I do not wish to compare these results, only to demonstrate with one case. In this sample the non-response ratio is 10% (which is considered to be very favourable in practice). I demonstrate the beneficial effect of imputation on this sample for the different methods.

I investigated the possible applications of classifying methods based on similarities in the identification of non-respondents. My investigation was based on cluster analysis, in which the first step is to find variables resulting in similarities and show a stochastic connection with the consumption of households but not correlating with each other.

Basic data for investigation:

- size of household (person)
- area of the property (m2)
- number of cars in the household
- number of television sets in the household
- employment status of the members of the household

Table 1 shows that I was able to identify the group of non-respondents using the above- listed variables, as 93.3 % of them belong to the cluster of better living conditions.

Table 1. Contingency table for grouping clusters and non-respondents

cluster membership	*measurements*	*respondents*	*non-respondents*	*total*
1st cluster: better circumstances	person percentage	531	84	615
2nd cluster: poorer circumstances	person percentage	65.6	93.3	68.3
total	person percentage	279	6	285

Proposition 3

In a case when non-respondents show similarities in material-type variables (consumption, income, etc.) then other features relating to the variable are examined and further similarities can be detected in the sample person in in terms of their demographic, social and economic factors, whichs make it possible to use classification methods to decrease non-sampling errors.

In addition to the use of cluster analysis results I also applied regression-based hot deck imputation.

In order to analyse the data in Table 2 we should note that the population parameter I tried to estimate is known: 1,744,633 Ft. It can be seen that from the entire population the population parameter can be estimated properly and with relatively low standard error. In cases when there is a data shortage the estimated value was significantly underestimated (by 9 %) compared to the population value.

The estimates deriving from samples resulting from imputations have improved compared to the data-shortage sample, as they mostly estimate a 4.5% lower value compared to the population parameter. Bias is still present in the estimates, as confidence intervals calculated on 95% reliability level do not meet the population parameter value. In this sense imputation methods cannot be considered successful, as they significantly underestimate the parameter.

Table 2. Estimated average household consumption expenditure

method	*estimate*	*standard error*	*coefficient of variation*	*relative difference from population parameter*	*amount of deviation (%)*	*confidence: (0: yes) (1: no)*	*relative standard error*
entire sample	1728151.47	8721.276	0.005	99.055%	0.945%	0	0.4999%
10% non response from upper 30%	1584815.61	7964.054	0.005	90.839%	9.161%	1	0.4565%
imputed iteration 5	1668815.81	12595.015	0.008	95.654%	4.346%	1	0.7219%
imputed from cluster with R estimate	1665379.74	10121.469	0.006	95.457%	4.543%	1	0.5801%
imputed with EV estimate	1668897.02	31730.657	0.019	95.659%	4.341%	1	1.8188%
imputed from cluster with EV estimate	1665488.15	31896.867	0.019	95.464%	4.536%	1	1.8283%
imputed upper 10% of cluster with EV estimate	1716799.67	33390.132	0.019	98.405%	1.595%	0	1.9139%
imputed upper 10% of cluster with R estimate	1716685.51	8618.436	0.005	98.398%	1.602%	1	0.4940%
Imputed Mahalanobis disntance	1620586.09	10633.907	0.007	92.890%	7.11%	1	0.6095%

Proposition 4

In a case of non-response simple imputation cannot reproduce the features of the original, entire sample, so it is very unlikely to make a reliable estimate, especially if the distribution of the population is asymetric. It can be sated that in a case when we estimate the linear statistics of a left-side asymetric population, then simple imputed estimates underestimate the population parameter with a significant bias.

Based on the statements of proposition 4, in addition to imputation we need other ways to reduce bias caused by non-response. As the quality of the sampling plan has an effect on the precision, reliability and efficiency, we may presume that a sampling that is carefully planned will reduce the bias in estimates caused by non-response.

In order to check the proposition I used the MR_Fogy_900 sample of consumption expenses, which shows the smallest relative error and the best sampling-plan effect. My main aim was that the stratum-forming criterion should have as close a stochastic connection as possible with the criterion used in the estimate. In my

case this expectation is fulfilled automatically as r=0.903**, showing a significant correlation between the two criteria.

After checking the conditions of the analysis I carried out the generation of non-responses, given the presupposition that questions concerning income and consumption are typically omitted by high-criterion value respondents. I chose the data-shortage items from a sample ranked depending on the consumption-expenses variable. The starting ratio was a very favourable 10% non-response rate. Afterwards I gradually increased this ratio by 5 % in each step up to the ratio of 50% non-responses.

At different response stages the artifically stratified sample resulted in the smallest estimates, in other words it contained the greatest bias.

Proposition 5(a)

In that case when the real or presumed reason for non-response shows a stochastic correlation with the stratum-forming criterion, stratification increases the scope of bias caused by non-response.

The explanation for the deviances from the result I expected could be found by analysing the dispersion. I concluded that for different sampling plans, the biggest differences occured in outer–deviance quadrat sums. In a case when entire group answers are missing, the widerdispersion of the other groups moderates the negative bias.

Proposition 5(b)

When researchers plan sampling they have to take the non-response expectations into consideration; if they potentially concern entire strata then it is worth modifying the representativeness requirements.

There are several methods for estimating non-response but I wanted to choose the one which requires the least possible outside information, as it is often not available to researchers. Foster (1996) experiments with several methods and concludes that the most successful results were achieved when they used variables from a census or microcensus.

In this dissertation I would like to reveal ways which provide opportunities to make proper estimates for enterprises and researchers hired by them.

On estimating the logistic regression-function I used different variable combinations and the composition of explanatory variables was the following:

- HD14_02: number of cars in the household,

– HA09: population density of the settlement,
– HC08: highest level of education completed,
– Income: income category.

With the parameters of regression function, I convert the estimated possible probabilities (Varga, 1999) into sample weights in order that potential non-respondent households would get more weight in estimating consumption expenses. This weighting system can reduce the scope of underestimation at different non-response levels, as is shown in the following table.

Table 3. Estimated average consumption expenses (in forint) for different levels of non-reponse, unweighted and weighted

level of non-response	*unweighted mean consumption expense*		*weighted mean consumption expense*		*relative difference of averages (unweighted=100%)*
	HUF	*percentage of expected value*	*HUF*	*percentage of expected value*	
upper 10% TC*	1,475,398	84.57%	1,554,136	89.08%	105.3%
upper 15% TC*	1,389,274	79.63%	1,494,852	85.68%	107.6%
upper 20% TC*	1,316,725	75.47%	1,428,680	81.89%	108.5%
upper 25% TC*	1,253,037	71.82%	1,371,626	78.62%	109.5%
upper 30% TC*	1,193,976	68.44%	1,316,734	75.47%	110.3%
upper 35% TC*	1,138,382	65.25%	1,266,237	72.58%	111.2%
upper 40% TC*	1,084,923	62.19%	1,214,319	69.60%	111.9%
upper 45% TC*	1,032,088	59.16%	1,172,525	67.21%	113.6%
upper 50% TC*	979,840	56.16%	1,105,332	63.36%	112.8%

* TC – total consumption

In the last column of the table you can see that the weighting improved the bias caused by non-response (by at least 5%). As the level of non-response increases the improving effect of weighting becomes better, which confirms the application of this method and that of the model. Howver, we cannot forget that practical experiments aim not only at reducing the negative effects of bias but also at estimating a population parameter as precisely and reliably as possible. This objective has only been partly achieved. From the population parameter we can see that a systematic increase in the level of non-response leads to a drastic increase in bias as well, despite the weighting.

Proposition 6

Weighting based on estimating non-response probability can reduce bias caused by non-response, but in the case of a systematic increase in non-responses the reducing effect of weighting is significantly less than the degree of bias.

It is obvious that application of complementary information helps with reducing the degree of error. Researchers do not always have the opportunity to use such information, so inner (within-sample) information should be utilised as much as possible. The tendency in responding groups (in the case of detailed grouping of sample items) can be projected to the entire sample, to the non-respondents. By modelling the tendencies we can reduce the bias caused by non-response.

In my study I formed groups based on household income to estimate the consumption expenses. I identified the exponential tendencies of consumption expenses in households with different incomes, paying attention to the different scope of non-response. I experienced that at lower levels of non-response the explanatory feature of functions was better; however, at higher levels the average of consumption expenses was considerably underestimated. That was the reason why I created an estimating model of weighted tendencies, in which the estimated values of the above-mentioned tendencies were defined as average estimated values by weighting the explanatory features of functions.

The model can be used at different non-response levels. In addition to the actual response ratio, the researcher has to generate non-responses in the groups artificially. The application of the model and the estimation procedure in the case of a 30% non-response rate are shown in the appendix.

Proposition 7

By generating artifical non-response levels, using the tendencies of respondents with the help of the estimate model of weighted tendencies, the population parameter can be estimated asymptotically.

The estimate model of weighted tendencies can be applied under the following conditions:

- there are criteria that determine non-response
- based on these criteria the population can be grouped into groups (preferably of the same size)
- there exists a mathematical function that describes the group tendencies significantly and reliably
- response ratio is larger than 50%.

If these conditions are met the model can provide a relatively good approximate value of average consumption expenses. Its greatest asset is that it can compensate for the significant underestimation which occured when using the imputation and transweighting methods.

The reducing effect of the model on bias caused by non-response is shown in Table 4.

Table 4. Comparison of results from non-response compensation methods by population parameter

level of non-response	*unweighted mean consumption expense*		*weighted mean consumption expense*		*estimate model of weighted tendencies*	
	HUF	*percentage of expected value*	*HUF*	*percentage of expected value*	*HUF*	*percentage of expected value*
upper 10% TC*	1,475,398	84.57%	1,554,136	89.08%	1,684,073	96.53%
upper 15% TC*	1,389,274	79.63%	1,494,852	85.68%	1,719,168	98.54%
upper 20% TC*	1,316,725	75.47%	1,428,680	81.89%	1,721,414	98.67%
upper 25% TC*	1,253,037	71.82%	1,371,626	78.62%	1,762,698	101.04%
upper 30% TC*	1,193,976	68.44%	1,316,734	75.47%	1,767,559	101.31%
upper 35% TC*	1,138,382	65.25%	1,266,237	72.58%	1,850,261	106.05%
upper 40% TC*	1,084,923	62.19%	1,214,319	69.60%	1,858,925	106.55%
upper 45% TC*	1,032,088	59.16%	1,172,525	67.21%	1,806,398	103.54%
upper 50% TC*	979,840	56.16%	1,105,332	63.36%	1,826.144	104.67%

* TC – total consumption

The only difference from Table 3 is in the last column headline: estimate model of weighted tendencies

Average consumption expenses counted with weighting or omitting non-responses can show as much as a 40% bias at higher non-response levels. At the same time the estimate model of weighted tendencies shows only a 10% bias. Nevertheless, the combined application of different methods is advisable, as we must bear in mind that the given sample is only one variation of the sampling plan, the examined criterion is a probability variable that can be influenced by several factors, and the models applied separately cannot solve every problem that might occur.

4. FIELDS OF APPLICATION FOR RESEARCH RESULTS

Research and conclusions based on sampling are gaining an ever larger role in both economic decision-making and the provision of information. The spread of sampling surveys can be attributed mainly to the reduced time and costs involved. Surveys based on sampling are becoming more popular not on a micro-level but also in macroeconomics. However, there is a major risk to this, due to the quality of samples.

For this reason, I believe that the findings and the methods developed in this study can be applied widely – in academic theory or by enterprises in their operation – in order to reduce the errors resulting from sample-based analyses and conclusions.

Data collection, evaluation and drawing appropriate conclusions are essential parts of research. If researchers take my models into account, these can contribute to the formulation of accurate data-based statements.

As far as education is concerned, I find it important to familiarize students not only with the various statistical procedures, relying on the above-mentioned methods, but also to teach them how to utilize more broadly the information hidden in sample data. I hope that my results will encourage students and young researchers to make efforts to reduce the degree of error found in data, rather than to simply accept it.

My experience shows that sometimes even public opinion polling organizations lack the information and methods which would enable them to properly process and evaluate the data obtained. It even occurs rather often that problems arise in the data intake which make the avoidance of exaggerated or mistaken conclusions nearly impossible. Thus I believe that institutions and companies working in these fields would be able to carry out their work more effectively if they considered my proposals when training poll-takers, compiling questionnaires, and evaluating the data.

Wide application of these results is ensured by the fact that the procedures used are applicable without substantial extra expenditure not only in compiling official statistics but also in company practice, and provide a basis for drawing correct conclusions.

BIBLIOGRAPHY

Ay János – Vita László: Egy kísérleti jövedelmi felvétel főbb tapasztalatai; Statisztikai Szemle, 1998. 76. évf. 6. szám, pp. 515-532.

Estevao V. M. – Särndal C. E.: Borrowing strength is not the best technique within a wide class of design-consistent domain estimators; Journal of Official Statistics, Vol. 20, No. 4, 2004, pp. 645–669

Estevao V. M. – Särndal C. E.: The ten cases of auxiliary information for calibration in two-phase sampling; Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 233–255.

Foster, K.: Weighting the Family Expenditure Survey in Great Britain to compensate for non-response: an investigation using census-linked data. Helsinki. 1996.

György Erika: A nemválaszolás elemzése a munkaerő-felvételben; Statisztikai Szemle, 82. évf. 2004. 8. sz. pp. 747-772

Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Havasi Éva – Schnell Lászlóné: Az 1996-os jövedelmi felvételre nem válaszoló háztartások – A megtagadások természete, a megtagadók sajátosságai; Központi Statisztikai Hivatal. Budapest. 1996.

Havasi Éva: Válaszmegtagadó háztartások; Statisztikai Szemle 1997. 10 sz. pp. 831-843.

Johansson F. – Klevmarken A.: Explaining the size and nature of response in a survey on health status and ecinimic standard; Journal of Official Statistics, Vol. 24 ,No. 3, 2008. pp. 431-449.

Keszthelyiné Rédei Mária: A lakossági jövedelmek mérésének megbízhatóbb módszere; Statisztikai Szemle, 2006. 84. évf. 5-6. szám pp. 518-551.

Kish, L: Kutatások statisztikai tervezése; Budapest, 1989.

KSH (1997): A háztartási költségvetési felvétel módszertana; Módszertani Füzetek 37. sz. KSH, Budapest.

KSH: A háztartások fogyasztása 2006, KSH, 2007.

Marton Ádám: A reprezentatív felvételek megbízhatósága 1991.

Roy D. – Safiquzzaman Md.: Variance estimation by Jackknife method under two-phase complex survey design; Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 35–51.

Sajtos László – Mitev Ariel: SPSS kutatási és adatelemzési kézikönyv; Alinea Kiadó, Budapest, 2007.

Särndal C. E. – Lundström S.: Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator; Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 167–191.

Varga Sára: A jövedelemfelvétel hiányzó adatainak pótlása; Statisztikai Szemle 1999. 77. évf. 2-3. sz. pp. 112-130.

PUBLICATIONS OF THE AUTHOR ON THE TOPIC

Books and papers

Faktoranalízis In: Kvantitatív információképzési technikák Miskolci Egyetem, Elektronikus tananyag, 2011. (megjelenés alatt)

Klaszteranalízis In: Kvantitatív információképzési technikák Miskolci Egyetem, Elektronikus tananyag, 2011. (megjelenés alatt)

A minta jellemzői; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 26-33. ISBN 978-963-06-7100-2

Szilágyi R. – Domán Cs.: Az adathiány kezelése mintavételes felmérésekben; Erdei Ferenc V. Tudományos konferencia – „Globális kihívások, lokális megoldások„ Kecskeméti Főiskola Kertészeti Főiskolai Kar Kecskemét, 2009. pp. 75-80. ISBN978-963-7294-74-7

Hipotézisvizsgálat; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 53-80. ISBN 978-963-06-7100-2

Szilágyi R. –Domán Cs.: Kalibráció a statisztikai becslésekben; „Gazdaság és társadalom" Nemzetközi tudományos konferencia Nyugat-magyarországi Egyetem Közgazdaságtudományi Kar Sopron, 2009. november 3. ISBN 978-963-9871-30-4

Mintavételi eljárások; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 9-25. ISBN 978-963-06-7100-2

Statisztikai becslés; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 33-52. ISBN 978-963-06-7100-2

Szilágyi R.: The infiltration of the unfounded statistical information in the forming mechanism of competitiveness In.: XXII. microCAD International Scientific Conference 2009. Miskolc pp. 233-238. ISBN 978-963-661-881-0

Besenyei L. – Domán Cs. – Szilágyi R. – Varga B.: Faktoranalízis alkalmazásának lehetősége az innovációs potenciál mérése során; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 45-52, ISBN: 978-963-661-798-1

Grafikus ábrázolás; In.: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 58-73. ISBN 978-963-06-3135-8

Ismérvek közötti sztochasztikus kapcsolatok elemzése; In.: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 140-153. ISBN 978-963-06-3135-8

Besenyei L. – Domán Cs. – Szilágyi R. – Varga B.: Klaszteranalízis alkalmazásának lehetősége az innovációs potenciál mérése során; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 53-64, ISBN: 978-963-661-798-1

Szilágyi R.: Mintavételes eljárások; Oktatási segédlet Miskolci Egyetem, 2007.

Szilágyi R.: Pénzbeli ellátások beilleszkedési kölcsönhatásai; In:„Globális és hazai problémák tegnaptól holnapig", VI. Magyar (Jubileumi) Jövőkutatási Konferencia, 30 éves az MTA IX. Osztály Jövőkutatási Bizottsága, Konferenciakötet 2., Budapest, Arisztotelész Stúdium Bt., 2007, pp. 91-97, ISBN: 978-963-86670-8-3

Besenyei L. – Domán Cs. – Szilágyi R. – Varga B.: „Statisztikai mintaillesztés" program tervezése és megvalósítása; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 8-16, ISBN: 978-963-661-798-1

Szilágyi R.: Statisztika az üzleti életben In: Informatikai statisztikus és gazdasági tervező felsőfokú képzés II. kötet 6. fejezet HEFOP-3.2.2-P.-2004-10-0011-/1.0 sz. projekt, Miskolc, 2007.

Viszonyszámok; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 42-57. ISBN 978-963-06-3135-8

Szilágyi R.: Kontár statisztikák; In: Doktoranduszok Fóruma Gazdaságtudományi Kar Szekciókiadványa, Miskolc, Miskolci Egyetem Innovációs és Technológia Transzfer Centrum, 2006, pp. 168-172.

Szilágyi R.: Minőségügyi statisztika; Oktatási segédlet Miskolci Egyetem, 2006.

Conference

Szilágyi R.: A nemválaszolás torzításának becslése a mintavételes felmérésekben; „HITEL, VILÁG, STÁDIUM" Tudományos konferencia, Sopron 2010. november 3.

Szilágyi R.: Analysis of nonresponse; International Conference "Economic & Social Challenges and Problems, at The time of Crisis 2009" Faculty of Economy, University of Tirana, Albania, 2009.

Besenyei L. – Varga B. – Domán Cs. – Szilágyi R.: Az elemezhetőséget biztosító mintaillesztés megvalósítása, Innovációmenedzsment, Tudásteremtés – Tudástranszfer Konferencia, Miskolc, 2006. november 15-16.

APPENDIX

Estimate model of weighted tendencies

Income tenths	*consumption at full response*	*30%NR*	*35% NR*	*40% NR*	*45% NR*	*50% NR*	*average estimated function-value*	*Estimated means of tenths*
1	*650,298*	749,355	742,179	734,640	731,450	722,461	736,179	*650,298*
2	*916,414*	887,189	883,780	880,178	879,221	876,507	881,437	*916,414*
3	*1,170,972*	1,050,374	1,052,398	1,054,549	1,056,846	1,063,400	1,055,428	*1,170,972*
4	*1,418,208*	1,243,575	1,253,186	1,263,464	1,270,355	1,290,142	1,263,851	*1,418,208*
5	*1,374,019*	1,472,313	1,492,283	1,513,767	1,526,999	1,565,231	1,513,536	*1,374,019*
6	*1,739,427*	1,743,123	1,776,998	1,813,658	1,835,492	1,898,976	1,812,674	*1,739,427*
7	*1,944,533*	2,063,746	2,116,034	2,172,959	2,206,308	2,303,883	2,171,083	*1,944,533*
8	*2,214,489*	2,443,342	2,519,755	2,603,441	2,652,038	2,795,126	2,600,538	2,329,175
9	*2,475,128*	2,892,759	3,000,503	3,119,205	3,187,817	3,391,114	3,115,157	2,790,094
10	*3,291,167*	3,424,840	3,572,973	3,737,147	3,831,836	4,114,181	3,731,872	3,342,456
mean	1,719,465	1,797,062	1,841,009	1,889,301	1,917,836	2,002,102	1,888,176	1,767,559
$\overline{R^2}$		0.9317	0.9408	0.9167	0.9218	0.8856		
function weight	1	0.20269	0.20467	0.19943	0.20054	0.19266		