Construction of the optimal design for the sample survey of small-scale enterprises of the service sphere of Ukraine

Artem Shcherbina[1] and Oksana Honchar[2]

[1] National Taras Shevchenko University of Kyiv, Ukraine
e-mail: artshcherbina@gmail.com

[2] Science and Technical Complex for Statistical Research, Ukraine
e-mail: ohonchar@list.ru

1 Introduction

We need to construct a sample of enterprises of Ukraine. This will be an annual investigation, that is based on a special accounting form. In this form there are several fields, such as unique code, geographic region, type of activity, annual turnover, average annual number of workers and some others.

Now in service sphere there are more than 60000 enterprises and this number increases. Earlier we had a census every year, so we have enough auxiliary information. This will undoubtedly be useful for constructing good sample design.

The sample should contain less than 20% of all enterprises and provide precise estimates not only for the whole country, but also for several subdomains, namely:

- different geographic regions (27)
- different types of activities (19)
- their intersections

Thus we need a stratified design with all intersections as strata. In 2006 there were 485 nonempty strata. The largest strata contains 1797 enterprises and the average strata size was 132. The coefficient of variation of turnover in strata had a maximum of 9.04, with average value 2.97. Let us consider now some strata and find the best sample design for it.

2 Estimation

Suppose we have strata of N enterprises with values of turnovers t_i, $i = 1, ..., N$. We want to estimate the total of these values. For the most enterprises we know values of the turnovers for previous year. Also we know previous value of the total turnover t_{last}. If N is small (<5) then the best method is to observe them all. The total sample limit is 20%, so let's set $n = \max(5, 0.2N)$.

2.1 Direct estimator

We have seen yet, that most of strata are heterogeneous, so, simple random sampling will provide bad estimates. Fortunately, we have additional information — data from previous censuses. Exploration of this data proves that values of turnovers for the same enterprise for different years are close to each other. Regression analysis shows that in average the turnover of an enterprise increases by 20% each year. For example, if some enterprise was observed 3 years before, we can expect $(1 + 0.2)^3$ increase of its turnover. Of course, it is quite roughly, but enough for use. One exception is small enterprises. Their future turnover doesn't depend on the previous value. So, it is better to set predicting values for them to some constant (average) value.

Thus for each enterprise that has ever been observed we know the expected value of turnover. As our estimates of turnovers are pretty precise, sampling without replacement with unequal probabilities proportional to prediction will provide the best results. In this design inclusion probabilities of enterprises π_i should be proportional to expected turnovers. As the sample size is n the probabilities can be computed as follows:

$$\pi_i = \frac{nt_i}{\sum_{j=1}^{N} t_j}$$

Here can occur the situation, when some obtained probabilities exceed 1. Then those enterprises include in sample for certain and inclusion probabilities recompute without them with smaller sample size.

The Horvitz-Thompson estimator of the total value of turnover is (Horvitz and Thompson 1952):

$$\hat{t}_{HT} = \sum_{i=1}^{N} \frac{I_i t_i}{\pi_i}$$

where $I_i = 1$ if object i is in the sample, and 0 otherwise. The variance of the Horvitz-Thompson estimator is:

$$V(\hat{t}_{HT}) = \sum_{i=1}^{N}\sum_{k>i}^{N}(\pi_i\pi_k - \pi_{ik})\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k}\right)$$

with unbiased variance estimator:

$$\hat{V}(\hat{t}_{HT}) = \sum_{i=1}^{N}\sum_{k=i+1}^{N} I_i I_k(\pi_i\pi_k - \pi_{ik})\left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_k}{\pi_k}\right)$$

Here are some problems. At first, all second-order inclusion probabilities π_{ij} must be positive. Second, in some cases obtained variance estimate can be negative. To satisfy this two conditions special sampling designs are used. Look at Brewer (1983) and Deville (1998) for some of them.

Another task is to decide what to do with enterprises that have never been observed. Let N_{mis} is number of such enterprises. In this case we have no idea about their turnover. Since we cannot distinguish them one from another, we need to select simple random sample of them. It is hard to choose the proper size n_{mis}, but it will be enough to take at least 10 enterprises and no more than 20% of all from each stratum. Then usual estimators of the total value $\hat{t}_{mis}$ and its variance $\hat{V}(\hat{t}_{mis})$ can be used:

$$\hat{t}_{mis} = \frac{N_{mis}}{n_{mis}}\sum_{i=1}^{N_{mis}} I_i t_i$$

$$\hat{V}(\hat{t}_{mis}) = N_{mis}^2\left(1 - \frac{n_{mis}}{N_{mis}}\right)\frac{s_{mis}^2}{n_{mis}}$$

where s_{mis} is sample variance:

$$s_{mis}^2 = \frac{1}{n_{mis} - 1}\sum_{i=1}^{N_{mis}} I_i(t_i - \bar{y}_{mis})^2$$

and $\bar{y}_{mis}$ — sample mean:

$$\bar{y}_{mis} = \frac{1}{n_{mis}}\sum_{i=1}^{N_{mis}} I_i t_i$$

The final direct estimate of the total value of turnovers for the strata is:

$$\hat{t}_{dir} = \hat{t}_{HT} + \hat{t}_{mis}$$

Since the previous two samples have been selected independently, the estimate of the variance of $\hat{t}_F$ is:

$$\hat{V}(\hat{t}_{dir}) = \hat{V}(\hat{t}_{HT}) + \hat{V}(\hat{t}_{mis})$$

2.2 Estimator based on prediction

In order to improve direct estimates for each stratum, consider the long-time behaviour of the total values of turnovers. Analysis shows that for quite big strata (>10 enterprises) ratios of total turnovers for the consecutive years are close to each other for different years and strata. For the last two years the average ratio of the total turnovers was 1.32 with standard deviation near 0.3. All this shows that for estimation of the stratum total we can use estimator

$$\hat{t}_{pred} = 1.32 * t_{last}$$

where $\hat{t}_{pred}$ is estimator based on prediction and t_{last} is the value of total turnover for the previous year. We don't know the true variance of this estimator, but we can expect that it will be $\hat{V}(\hat{t}_{pred}) = 0.3$.

2.3 Composite estimator

Thus we get two estimators: the direct estimator $\hat{t}_{dir}$ with variance $\hat{V}(\hat{t}_{dir})$ and the estimator based on prediction $\hat{t}_{pred}$ with assumed variance $\hat{V}(\hat{t}_{pred})$. Consider the composite estimator

$$\hat{t}_{comp} = b\hat{t}_{dir} + (1-b)\hat{t}_{pred}$$

As estimates of strata total are independent, the minimum mean square error of $\hat{t}_{comp}$ will be attained at

$$b^* = \frac{V_{pred}}{V_{dir} + V_{pred}}$$

with variance

$$\hat{V}(\hat{t}_{comp}) = \frac{V_{dir} * V_{pred}}{V_{dir} + V_{pred}}$$

The procedure of constructing composite estimators is explained by Longford (2005).

References

Bethlehem, J. G. & Schuerhoff, M. H. (1984) *Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities.* Biometrika 71, 653-6

Brewer, K. R. W. & Hanif, M. (1983). *Sampling with Unequal probabilities.* New York: Springer-Verlag.

Deville, J.-C. & Tillé, Y. (1998). *Unequal probability sampling without replacement through a splitting method.* Biometrika 85, 89-101.

Lohr. S.L. (1999). *Sampling: design and analysis.* Duxburry Press.

Longford, N. T. (2005). *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician.*

Sampford, M. R. (1967). *On sampling without replacement with unequal probabilities of selection.* Biometricka 54, 499-513.

Yates, F. & Grundy, P.M. (1953). *Selection without replacement from within strata with probability proportional to size.* Journal of the Royal Statistical Society, B 15, 235-61.