FROM THE EDITOR

This issue consists of eleven articles, a report from the didactic conference in Poland, and an announcement on the 8th Conference of the International Federation of Classification Societies. Out of eleven articles six concern Sampling Methods, one deals with Statistical Education and further four papers can be found in section Other Articles.

The articles on sampling methods form the biggest subject group and include:

1. Sample Designs for National Surveys: Surveying Small-Scale Economic Units (by Vijay Verma, Professore a Contratto, Università di Siena, Director, ORC Macro International Social Research, London). The author discusses some special considerations which arise in the design of samples for what may be termed economic surveys, as distinct from population-based surveys. By this he means surveys concerned with the study of characteristics of economic units, such as agricultural holdings, household enterprises, own-account businesses, or other establishments in different sectors of the economy. His focus is exclusively on sample surveys of economic units which, like households, are small-scale, numerous and widely dispersed in the population. Units which are medium-to-large in size, few in number or are not widely dispersed may require different approaches, often based on list frames.

2. Influence of Numbers of Grouped Balanced Half-Samples on Effectiveness of Variance Estimation for Complex Sample Surveys (by J. Jakubowski and Cz. Bracha from Poland). In this paper the authors consider the problem of approximate variance estimation in survey sampling. They investigate repeatedly grouped balanced half-sample method (Rao-Shao method). By means of simulation study, the complex sampling survey has been examined. Results confirm the usefulness of Rao-Shao method in the practice of survey sampling.

3. On the Use of Transformed Auxiliary Variable in the Estimation of Population Mean in Two Phase Sampling (by G.N. Singh from India). The paper deals with the problem of estimating the population mean in two-phase sampling. New chain type estimators have been proposed. They make use of the known standard deviation, coefficient of skewness and coefficient of kurtosis of the second auxiliary character. The proposed estimators have been
compared with two-phase ratio estimator and some other chain type estimators. Further generalisations have been made and their properties are discussed. The performances of the proposed estimators have been supported with a numerical illustration.

4. **On Some Equivalent Definitions of Simple Random Sampling and Theoretical Justification of Various Selection Methods Thereby** (by G.C. Tikkiwal from India). Tikkiwal, B.D. (1984) gives three equivalent definitions for each of the two sampling schemes: the Simple Random Sampling with and without replacement. This paper gives a rigorous proof of the equivalence of these definitions in each case. Making use of these three equivalent definitions, it then gives theoretical justification of the selection of simple random samples through various methods proposed in the literature.

5. **The Use of a Known Coefficient of Variation in the Estimation of Mean of a Normal Distribution from Double Samples** (by L. N. Upadhyaya and S. R. Srivastava from India). In this paper the authors have proposed an estimator of the population mean $\mu$ from double samples when the coefficient of variation of the normal distribution is known and they have a guessed value of the population mean $\mu$. The properties of the proposed estimator have been discussed. Numerical values of the relative efficiencies have been obtained and recommendations regarding its use have been made.

6. **Estimation of the Population Mean on the Basis of Non-Simple Sample When Non-Response Error is Present** (by J. Wywial from Poland). The entire sampling design and sampling scheme are determined by the appropriate set of the probabilities of inclusion of the first and second order. The estimator of a population mean is considered in the case when a non-response error is present.

One article concerns **Statistical Education**:

7. **Some Aspects of Statistical Training in Banking in Ukraine** (by V S. Gerasymenko and N. Golovach from Ukraine). This article follows a series of publications about teaching applied statistics in Ukraine (see: Statistics in Transition, vol. 3, Number 3, June 1998). As it is known, the bank system both produces and consumes a large volume of information. Its collection, analysis and use of the results of the analysis for effective management of bank activities requires statistical knowledge. Bank Statistics, a new applied statistics course, has been taught at bank schools in Ukraine for 5 years. Since 2001 Bank Statistics has been taught as a part of special programs in all universities.

There are four articles in section **Other articles**:

8. **Technological Level and the Capacity to Innovate of Small Enterprises in Poland** (by A. Balicki from Poland). The position of small enterprises and
their role in the coming years is a combined result of numerous, usually unfavourable factors of both macroeconomic and internal character. The question is whether the importance of the small enterprise sector will increase, as it is expected by theoreticians, or it will be rather observed a great number of failures of companies. In this paper the author intents to answer the question about the extent to which the small businesses have the capacity to develop and whether their owners are sufficiently entrepreneurial to overcome difficult situations and survive the increasing competition.

9. Some Data Quality Issues in Statistical Publications in Poland (by J. Kordos from Poland). This paper focuses on the reporting and presentation of information on sources of error in several dissemination media: short-format publications, main report from the surveys, analytic publications, and the Internet. This review is based on the work of the author conducted for training purposes and in order to help characterise current practices for reporting sources of error in publications. The first study reviewed press releases and short publications of several pages (“short-format reports) issued by six divisions of the Central Statistical Office (CSO): (i) labour, (ii) demography (iii) living conditions, (iv) agriculture, (v) enterprise, and (vi) trade &services. The second study reviewed main reports from the sample surveys published by the same divisions. These are printed reports of methodology of the survey, descriptive analysis, and tables with main results from the survey. The third study reviewed selected “analytic publications” from the same divisions. The fourth review is connected with the Internet which has become one of the media of dissemination of statistical data from the CSO since 1997.

10. Remarks to the Utilisation of the Concept of Linear Rectifications of Repetitive Forecasts (by J. Kozak from the Czech Republic). When statistical forecasts based on the analysis of a time-series of the indicators are constructed, one often faces a situation when building an appropriate model is questionable or sometimes even practically impossible. Such situations arise mostly when the time-series are of a relatively small size, "good" explanatory variables are absent, etc. Therefore some acceptable solutions of the forecasting problem in these situations are of practical interest.

11. Estimation of Finite Population Distribution Function Using Multivariate Auxiliary Information (M S Ahmed from Bangladesh and W. Abu-Dayyeh from Jordan). This paper derives the generalised estimator for the finite population distribution function using multivariate auxiliary information. The properties of this estimator are given for the general sampling design. The results for the simple random sampling without replacement are presented for the estimator. Finally, a simulation study is carried out from a real data set for the relative comparisons.
There is also one report on the *Tenth Didactical Conference on Methods of Qualitative Attributes Analysis in Decision Making Process* which was held in Lodz, Poland, from 4 to 5 June 2001; and an announcement on the *8th Conference of the International Federation of Classification Societies* to be held in Cracow, Poland, July 16-19, 2002.

Jan Kordos
The Editor
SAMPLE DESIGNS FOR NATIONAL SURVEYS: SURVEYING SMALL-SCALE ECONOMIC UNITS1

Vijay Verma2

ABSTRACT

The paper discusses special considerations which arise in the design of samples for economic surveys as distinct from household surveys, i.e. in surveys concerned with the study of characteristics of economic units (agricultural holdings, household enterprises, own-account businesses, or other establishments in different sectors of the economy). The units considered, like households, are small-scale, numerous and widely dispersed in the population, but differ in being much more heterogeneous and unevenly distributed. The central requirement of the design is the assignment of measures of size for the probability proportional to size (PPS) selection of area units in such a way that the sample size requirements in terms of the ultimate units (establishments) by economic sector are met. A useful technique is to define ‘strata of concentration’ classifying area units according to the predominant type(s) of establishments contained in the area, and use this structure to assign measures of size to control the unit probabilities of selection to achieve the required sample allocation. The paper develops various strategies of assigning measures of size to area units, and evaluates them in terms of the achieved efficiency of the design. This type of design has a wide variety of practical applications, an example being the selection of schools while controlling the allocation of the sample according to ethnic group of the student population.

Key words: sample design, small scale economic unit, agricultural holding, household enterprise, probability proportional to size, sample size requirement

1 An earlier version of this paper was presented as invited contribution at Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR), 5-7 June, 2001. It is intended to submit at a later date to Statistics in Focus some empirical results from the application of the methodology described herein.

2 Professore a Contratto, Università di Siena, Director, ORC Macro International Social Research, London, Angel Corner House, 1 Islington High Street, London N1 9AH United Kingdom. tel (44-20) 7675 1063; fax (44-20) 7675 1906; vijay.verma@orc.co.uk
1. Introduction

Much discussion of sampling methods, including in textbooks on the subject, tends to be confined to the design of population-based surveys, i.e. surveys in which households (or sometimes individual persons) form the ultimate units of selection, collection and analysis. The theory and practice of large-scale population-based sample surveys is reasonably well established and understood.

In this paper I will discuss some special considerations which arise in the design of samples for what may be termed economic surveys, as distinct from population-based surveys. By this I mean surveys concerned with the study of characteristics of economic units, such as agricultural holdings, household enterprises, own-account businesses, or other establishments in different sectors of the economy. My focus will be exclusively on sample surveys of economic units which, like households, are small-scale, numerous and widely dispersed in the population. Units which are medium-to-large in size, few in number or are not widely dispersed may require different approaches, often based on list frames.

The type of sample designs used in 'typical' household surveys provide the point of departure in this discussion of sampling of other small-scale units. Indeed, there may often be a one-to-one correspondence between such economic units and households, and households rather than the economic units themselves may directly serve as the ultimate sampling units. Nevertheless, despite much common ground with sampling for population-based household surveys, sampling small-scale economic units involves a number of different and additional considerations.

It is useful to begin by noting some similarities between the two situations, and then move on to identifying and developing special features of sampling for surveys of small-scale economic units.

2. Similarities with household survey design

National or otherwise large-scale household surveys are typically based on multi-stage sampling designs. Firstly, a sample of area units is selected in one or more stages, and at the last stage a sample of ultimate units (dwelling, households, persons etc) is selected within each sample area. Increasingly – including, and especially in, developing countries – a more or less standard two-stage design is becoming common. In this design the first stage consists of the selection of area units with probability proportional to some measure of size, \( M_k \), such as the estimated number of households or persons in area \( k \) from some past source providing such information for all areas in the sampling frame. At the second stage, ultimate units are selected within each sample area with probability inversely proportional to size. The overall probability of selection of a unit in area \( k \) is
\[ f_k = \left( \frac{aM_k}{M} \right) \left( \frac{b}{N_k} \right) = \frac{M_k}{N_k}, \quad (1) \]

where \((a, b, M, f)\) are constants. Here \(a\) is the number of areas selected, when \(M\) is the total of \(M_k\) values in the population; \(b\) is the expected number of ultimate units selected per sample area; hence \(a\times b = n\) is the expected sample size; and \(f\) is a constant defined as

\[ f = \frac{aM}{M} = \frac{n}{M}. \quad (2) \]

The denominator \(N_k\) may be the same as \(M_k\) (the measure of size used at the first stage), in which case we get a self-weighting sample with \(f_k = f = \text{const.}\) Alternatively, \(N_k\) may be the actual size of the area, in which case we get a 'constant take' design, i.e. with a constant number \((b)\) of ultimate units selected from each sample area irrespective of the size of the area. It is also possible to have \(N_k\) as some alternative measure of size, for instance representing a compromise between the above two designs. In any case, \(M_k\) and \(N_k\) are usually closely related to – and are meant to approximate – the actual size of the area.¹

It is common in national household surveys to aim at self-weighting or approximately self-weighting designs. This often applies at least within major geographical domains such as urban-rural or regions of the country.

The selection of ultimate units within each sample area requires a listing of these units. Often existing lists have to be updated or new lists prepared for the purpose to capture the current situation. No such lists are required for areas not selected at the first stage. The absence of up-to-date lists of ultimate units for the whole population is a major reason for using area-based multi-stage designs.

Now let us consider a survey of small-scale economic units such as agricultural holdings or other types of household enterprises in similar circumstances. Just like households, such units tend to be numerous and dispersed in the population. Indeed, households themselves may form the ultimate sampling units in such surveys, the economic units of interest coming into the sample through their association with households. Similar to the situation with household surveys, typically no up-to-date lists of small-scale economic units are available for the entire population. This requires resorting to an area-based multi-stage design, such as the one implied by (1) above.

The main difference is that economic surveys generally require major departures from self-weighting designs. We explore and discuss this fundamental difference in the remainder of this paper.

3. Special features of economic surveys

¹ Throughout, ‘size’ refers to the number of ultimate units in the area, not to its physical size.
Despite similarities noted above, there are certain major differences in the design requirements of population-based household surveys and surveys of small-scale (often household-based) economic units. These arise from differences in the type and distribution of the units and in the reporting requirements.

**Heterogeneity**

Household surveys are generally designed to cover the entire population uniformly. Different subgroups (such as households by size and type, age and sex groups in the population, social classes etc) are often important analysis and reporting categories, but (except possibly for geographical classes) are rarely distinct design domains. By contrast, economic units are characterised by their heterogeneity and by much more uneven spatial distribution. The population comprises of multiple 'sectors', often with great differences in the number, distribution, size and other characteristics of the units in different sectors – representing different types of economic activities to be captured in the survey, possibly using different questionnaires and even data collection methodologies. Separate and detailed reporting by sector tends to be a much more fundamental requirement, than it is in the case for different population subgroups in household surveys. The economic sectors can, and often do, differ greatly in size (number of units in the population) and in sample size (precision) requirements, and hence in the required sampling rates. Therefore it is necessary to treat them not only as separate analysis and reporting categories, but also as distinct design domains.

**Uneven distribution**

These aspects are accentuated by uneven geographical distribution of economic units of different types. Normally, different sectors to be covered in the same survey are distributed very differently across the population: varying from (i) some sectors concentrated in a few areas, to (ii) some widely dispersed throughout, but with (iii) many 'mixed' sectors concentrated or dispersed to varying degrees. These patterns of geographical distribution have to be captured in the sampling design. True, population subgroups of interest in household surveys can also differ in their distribution (as in the typology of 'geographical, 'cross' and 'mixed' subclasses defined by Leslie Kish in the analysis of design effects), but normally type (ii) rather than (iii) predominate there. By contrast, often situation (iii) predominates in economic surveys; and furthermore, as noted above, such 'mixed' sectors need to be treated as distinct design domains.

**Sampling versus survey units**

There are a number of other factors which make the design of economic surveys more complex than that of household surveys. Complexity arises from the possibility that the ultimate units used in sample selection may not be of the same type as the units involved in data collection and analysis. The two types of units may lack one-to-one correspondence. For instance, the ultimate sampling units may be (often are) households, each of which may represent none, one or more
than one type of economic activity of interest. For instance the same household may undertake different types of agricultural activities. Hence, seen in terms of the ultimate sampling units (households), different sectors (substantive domains) are not disjoint but overlapping. This gives rise to two possible design strategies: (1) An integrated design, based on a common sample of households, in which all sectors of activity in which a selected unit is engaged in would be covered simultaneously. (2) Separate sectoral designs, in which the sector populations (in terms of the sampled units, households) and hence the samples may overlap. In each sectoral survey, activity of the selected households pertaining only to the sector concerned would be enumerated.

Separate surveys are generally more costly and difficult to implement. The overlap between the sectoral samples may be removed by characterising the sampling units (households) in terms of their *predominating* sector. This helps to make the sampling process more manageable. However, this precludes separate sectoral surveys: in so far as the sample for any particular sector is restricted only to the households in which that sector predominates over all other sectors, the coverage of the sector remains incomplete.

**Sampling different types of units**

In practice it is often costly, difficult and error-prone to identify and separate out the ultimate survey units into different sectors and apply different sampling procedures or rates by sector. Hence it is desirable, as far as possible, to *absorb any differences in the sampling requirements by sector at preceding area stage(s) of sampling*, so as to avoid having to treat different types of units differently at the ultimate stage of sampling. Of course, the cost of such (very desirable) operational simplification is the increased complexity of the design this may involve.

**4. Description of the basic design**

Let us now consider some basic features of an area-based multi-stage sampling design for a survey of small-scale economic units. The population of units comprises a number of ‘sectors’, such as different types of holdings, agricultural activities or products. Sample size requirements have been specified for each sector. The available sampling frame consists of area units (which form the primary sampling units (PSU’s) in a two-stage design, or the 'ultimate area units' (UAA's) in a design with multiple area stages). Information is available on the expected number of economic units ($N_{i,k}$) in each area $k$ by sector $i$, and hence on their total ($N_k$) for the area.$^1$

---

$^1$ Note that this information requirement by sector is more elaborate than a single measure of population size normally required for probability proportional to size (PPS) sampling in a household survey. It is important that, especially in the context of developing countries with limited administrative sources, potential sources such as population, agricultural and economic
In essence, the overall selection equation (following equation [1] above) is of the form:

\[ f_k = f \left( \frac{M_k}{N_k} \right), \]  

(3)

where \( M_k \) is some measure of size assigned to the area in its PPS selection, and the selection of ultimate units within the area is with probability inversely proportional to the size measure \( N_k \) in the frame, assumed to estimate the actual size of the area.

The design weight to be applied at the estimation stage is inversely proportional to the overall selection probability:

\[ w_k = \frac{1}{f_k} \alpha \left( \frac{N_k}{M_k} \right). \]  

(4)

The expected number of units contributed to the sample by area \( k \) is

\[ f_k \cdot N_k = f \cdot M_k \]  

(5)

Note that the reference here is to the expected value of the contribution of the area to the sample. The actual number of units contributed by any area will be zero if the area is not selected at the preceding stage(s), and generally much larger if the area has been selected.

Summing over all areas \( k \) in the population in [5] gives the total expected sample size

\[ n = f^* M. \]  

(5a)

The expected number of units of a particular sector \( i \) contributed to the sample by area \( k \) is

\[ f_k \cdot N_{i,k} = f \cdot M_k \cdot \left( \frac{N_{i,k}}{N_k} \right) = f \cdot M_k \cdot P_{i,k}, \]  

(6)

The basic design problem is to determine the 'modified' size measures \( M_k \) such that the sample size requirements

\[ n_i = f \cdot \Sigma_k \left( M_k \cdot P_{i,k} \right), \]  

(7)

or in terms of relative quantities more convenient for numerical work

\[ \frac{n_i}{n} = \Sigma_k \left( \frac{M_k}{M} \right) P_{i,k}, \]  

(7a)

censuses are designed to yield such information, required for efficient design of surveys of small-scale economic units.
are satisfied for all sectors \(i\) simultaneously in the most efficient way. The criterion of 'efficiency' also need to be defined (see next section).

In [7] the sum is over all areas in the population (and not merely the sample). Note also that the above formulation assumes that at the ultimate sampling stage, units within a sample area are selected at a uniform rate, inversely proportional to \(N_k\), irrespective of the sector. This is a very desirable feature of the design in practice. As noted earlier, it is often costly, difficult and error-prone to identify and separate out the ultimate survey units into different sectors and apply different sampling procedures or rates by sector. The preceding sampling stages are assumed to absorb any difference in the required sampling rates by sector through incorporating those in the definition of the size measures \(M_{k(i)}\), yet to be defined.

The most convenient (but also the most unlikely) situation in the application of [7] is when units of different types (sectors) are geographically completely segregated, i.e. when each area contains units belonging to only one particular sector. Indicating by \(k(i)\) areas in the set containing units of sector \(i\) only (and of no other sector), it can be seen that [7] is satisfied by a simple relationship of the form

\[
M_{k(i)} = g_i \cdot N_{k(i)},
\]

(8)
giving from (7)

\[
N_i = f \cdot g_i \cdot \sum_{k(i)} N_{k(i)} = f \cdot g_i \cdot N_i
\]

or

\[
g_i = \left( \frac{N_i}{N_i} \right) = \left( \frac{f_i}{f} \right),
\]

(9)
i.e., inflation of all size measures \(N_{k(i)}\) in proportion to the required overall sampling rate for the sector.

It is useful to note that once the size measures are so inflated, the required differences in the sectoral sampling rates are automatically ensured and it is no longer necessary to apply the sample selection operation separately by sector. The use of adjusting size measures to simplify the selection operation is a useful and convenient device, applicable widely in sampling practice.

In reality the situation is more complex because areas generally contain a mixture of units of different types (sectors), and a simple equation like (8) cannot be applied. Clearly, we should inflate the size measure (and hence inflate the selection probabilities) for areas with proportionately more units from sectors which need to be over-sampled, and vice versa. These considerations need to be quantified more precisely. We know of no exact or theoretical solutions to [7], and have to seek empirical (numerical) solutions determining \(M_k\) for the PPS selection of areas, solutions which involve trial and error and defy strict optimisation.
5. Evaluation criterion: the effect of weights on sampling precision

The effect of ‘random’ weights

The design effect, which measures the efficiency of a sample design compared to a simple random sample of the same size, can be decomposed under certain assumptions into two factors:

- the effect of sample weights, and
- the effect of other aspects of the sample design, such as clustering and stratification.

We are concerned here with the first component – the effect of sample weights on precision. This effect is generally to inflate variances and reduce the overall efficiency of the design. This arises from the difference between the actual size measures $N_k$, and the modified size measures $M_k$ used in the selection of area units with the objective of meeting the sample size requirements $n_i$ by sector. The increase in variance depends on the variability in the selection probabilities or the resulting design weights (equation [4]). We use the following equations to compare different choices of the modified size measures in terms of their effect on efficiency of the resulting sample in an empirical search for the best, or at least a 'good', solution.

Such decomposition is possible when, as in the present case, weights are 'external' or 'arbitrary', i.e. essentially uncorrelated with population variances. It has been established theoretically and empirically that the effect of such weighting tends to persist uniformly across estimates for diverse variables and population subclasses, including estimates of differentials and trends, and is well approximated by the following expression:

$$D^2 = (1 + cv^2(w_j)), \quad (10)$$

where $cv(w_j)$ is the coefficient of variation of the weights of the ultimate units in the sample. The expression approximates the factor by which sampling variances are inflated, i.e. the effective sample size is reduced.

From weights $w_j$ for individual units $j$ in the sample of size $n$, the above can be written as follows, with $\Sigma$ representing the sum over units in the sample:

$$D^2 = \left( \frac{\sum w_j^2}{n} \right) / \left( \frac{\sum w_j}{n} \right)^2, \quad (11)$$

or, for sets of $n_k$ units with the same uniform weight $w_k$ the above becomes:

$$D^2 = \left( \frac{\sum n_k \cdot w_k^2}{\sum n_k} \right) / \left( \frac{\sum n_k \cdot w_k}{\sum n_k} \right)^2. \quad (12)$$
**Computation of $D^2$ from the frame**

It is useful to write the above equations in terms of weights of units in the population, so that different design strategies can be evaluated without actually having to draw different samples:

$$D^2 = \left( \frac{\sum\left(1/w_j\right)}{N} \right) \left( \frac{\sum(w_j)}{N} \right),$$

or, for sets of $N_k$ units with the same uniform weight $w_k$ the above becomes:

$$D_i^2 = \left( \frac{\sum\left(N_{i,k}/w_k\right)}{\sum N_{i,k}} \right) \left( \frac{\sum\left(N_{i,k,w_k}\right)}{\sum N_{i,k}} \right), \quad w_k = \frac{N_k}{M_k}$$

The above equation applies to the total population, as well as separately to each sector $i$. Ideally, subsampling within any area $k$ is identical for all sectors ($i$), implying uniform weights $w_k$ for all types of units in the area. These weights are proportional to $(N_k/M_k)$ for the area concerned.

The loss due to weighting also has an effect on the effective allocation by sector actually achieved. The effective sample size $n'_i$, in the presence of arbitrary weights is smaller by the factor $(1/D_i^2)$, compared to the actual sample size $n_i$. The average of $D_i^2$ values over the sectors

$$\frac{\sum_i D_i^2}{I_i}$$

may be taken as an overall indicator of the inflation in variance due to weighting in the sectoral designs. The objective is to minimise this indicator by appropriately choosing the size measures $M_k$ satisfying the required sample allocation (7).

**6. Constructing "strata of concentration" (StrCon)**

To adjust the measures of size $M_k$ and hence the overall sampling rates to achieve the required sample size by sector, it is useful to begin by classifying areas into groups according to which particular sector predominates in the area. The basic idea is as follows.

For each sector, the corresponding 'stratum of concentration' is defined to consist of a set of areas in which that sector 'predominates' in the sense as defined below. One such stratum corresponds to each sector. The objective of constructing such strata is to separate out areas according to their composition in terms of sector. (To distinguish these from 'ordinary' strata used for sample selection, I will henceforth refer to them as "StrCon").
**Notation**

i subscript identifying a sector, i=1 to I.

k subscript identifying an area

$N_{i,k}$ number of units of sector i in area k

$N_k$ total number of units in area k (all sector combined)

$N_i$ total number of units of sector i (all areas in the population)

$A_i$ number of areas containing at least 1 units of sector i ($N_{i,k}>0$)

$B_i$ average number of units of sector i per area (counting only areas containing at least 1 such unit) = ($N_i$/ $A_i$)

$P_{i,k}$ 'index of relative concentration' of sector i in area k = $N_{i,k}$ / $B_i$

j index (1-3) identifying the stratum of concentration (StrCon), i.e. the sector i which has the highest $P_{i,k}$ value for the area k

Note that the 'index of relative concentration' $P_{i,k}$ has been defined in relative terms: the number of units of a particular sector i in area k in relation to the average number of such units per area. Defining this simply in terms of the actual number of units will result in automatic over-domination of the largest sectors. However, for a sector not large in overall size but confined to only a small proportion of the areas, the average per area (including zeros) would tend to be small, resulting in its over-domination. Hence it appears appropriate to exclude zeros (areas with $N_{i,k}=0$) in computing the average $B_i$.

Data by StrCon and sector (aggregated over areas).

Once the StrCon has been defined for each area, the basic information aggregated over areas on the numbers of units classified by sector and StrCon can be represented in a square table of the following form (Table 1).

<table>
<thead>
<tr>
<th>StrCon</th>
<th>1</th>
<th>...</th>
<th>j</th>
<th>...</th>
<th>I</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sector</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nij</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>N.j</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following notation is used in the table.
Nij refers to the number of units of sector i in all areas in StrCon j. Summing along rows,
Ni. is the total number of units in sector i, \( \sum_j N_{ij} \). Summing along columns,
Nj. is the number of units in StrCon j, \( \sum_i N_{ij} \).
ni. is the target sample size for sector i, 
fi = (ni. / Ni.) is the implied overall sampling rate for sector i.
Nk(j) where necessary, k(j) will be used to refer to a particular area k in StrCon Hence for instance, summed over areas in the frame for StrCon (j): 
\( N_{ij} = \sum_k N_{i,j(k)} \), \( N_{j.} = \sum_k N_{j,k} \).

7. Using StrCon for determining the sampling rates

A basic model

The use of StrCon as defined above is the fundamental aspect of the strategy for assigning measures of size to areas for the purpose of achieving the required sample allocation by sector.

We will first consider a simple model using a uniform sampling rate within each StrCon, but varied across StrCon with the objective of obtaining the required allocation by sector. This basic model will be refined subsequently.

The model implies the assignment of area measures of size in the form

\[
M_{k(j)} = g_{j}.N_{k(j)}, \tag{16}
\]

where the actual size \( N \) of area k in StrCon j is inflated by a factor \( g_j \) constant for the StrCon. This is the factor by which the ultimate selection probabilities of units in the area are inflated. Note that in [16] the scale of \( M \) and hence of \( g_j \) is arbitrary. For numerical convenience, it can be chosen such that for [16] summed over the whole frame, \( M = N \).

Since the above inflation in this basic model applies equally to units of all sectors in an area, it can be written in terms of cells of Table 1 as

\[
M_{ij} = g_{j}.N_{ij}. \tag{17}
\]

Equation [16] looks formally the same as [8]; but the latter is a particular and very simple case of [16], for the situation when Table 1 has non-zero elements only in the diagonal (representing complete geographical separation of different sectors).

Suppose that the objective is to obtain sample size \( n_i \) for sector i, i.e. to apply the average sampling rate

\[
f_i = (n_i / N_{i.}). \tag{18}
\]

This rate cannot, of course, be applied directly to rows of Table 1 (i.e. to numbers of units by sector), but must be applied to (whole) areas, i.e. along columns of the table. The basic requirement is to solve the following set of linear equations to determine the StrCon sampling rates say \( (f_j. g_j) \) - or its exact equivalent, the
required adjustment to measures of size [10] - given the distribution $N_{ij}$ and the
target sample sizes $n_i$ by sector $i$:

$$f \cdot \sum_j \left(M_{ij}ight) = f \cdot \sum_j \left(g_j \cdot N_{ij}ight) = f_i \cdot N_i = n_i, \text{ for } j, i = 1 \text{ to } I. \quad (19)$$

or in terms of relative quantities more convenient for numerical work

$$\frac{n_i}{n} = \sum_j g_j \left(\frac{N_{ij}}{N}\right). \quad (19)$$

In principle, the above can be solved by inverting an $(I \times I)$ matrix. However,
an iterative solution can be simpler and more convenient. To the extent diagonal
elements predominate in Table 1, a good starting point is to assume $g_j = \frac{f_i}{f}$ for
$j=i$, and then iteratively adjust the $g_j$ values to satisfy \[19\] within a certain margin.

**More flexible models: an empirical approach**

Depending on the numbers and distribution of units of different types and on
the extent to which the required sampling rates by sector differ, a basic model like
[16] may be too inflexible, and may result in large variations in design weights
and hence large losses in efficiency of the design. It may even prove impossible to
satisfy the sample allocation requirements.

Lacking a general theoretical solution, we have tried more flexible empirical
approaches to defining the modified size measures to meet the required sample
allocation and to achieve this more efficiently.

Basically, the approach has involved supplementing [16] by further
modifying the size measures in a more targeted fashion, as follows.

The original size measure broken down by sector is

$$N_{\ast j(i)} = \sum_i N_{i,\ast j(i)} = (N_{\ast k(j)} - N_{i,\ast k(j)}) + (N_{j,\ast k(j)}), \quad (20)$$

and the modified measure is defined in the form

$$M_{\ast j(i)} = g_j \left(\frac{N_{\ast k(j)} - N_{i,\ast k(j)}}{N_{j,\ast k(j)}} + h_{j,\ast k(j)} \right), \quad (21)$$

The first term on the right in [20] is the original size measure in terms of all
units other than those belonging to the sector ($i=j$) corresponding to the area's StrCon ($j$). The second term is the original size measure for units belonging to that sector. It is the latter which is modified by some factor $h$ in [21] to facilitate
meeting the sample allocation constraints. The variation in the overall sampling
rates is determined by the ratio (21)/(20):

$$\frac{M_{\ast j(i)}}{N_{\ast j(i)}} = g_j \left[(1-x) + h_j (x) x\right], \quad (22)$$
where 'x' stands, for a given area \( k \), for the proportion of units belonging to the sector \((i=j)\) which corresponds to StrCon \((j)\) of the area:

\[
\frac{N_{j,k(j)}}{N_{j,k(j)}} = P_{j,k(j)} = x, \text{ say.} \tag{23}
\]

It is assumed in [22] that an appropriate form for \( h_j \) is to take it as a function of \( x \); subscript \( j \) indicates that form may differ by StrCon.

The above can be aggregated over all areas \( k(j) \) within each StrCon to construct cells of Table 1:

\[
N_j = \sum_{k(j)} N_{j,k(j)},
\]

\[
M_j = \sum_{k(j)} \left( \frac{M_{j,k(j)}}{N_{j,k(j)}} \right) N_{j,k(j)} = g_j \sum_{k(j)} \left( (1-x) + h_j(x) x \right) N_{j,k(j)} = g_j N'_j, \text{ say.} \tag{24}
\]

The sample allocation constraint is

\[
f \cdot \sum_j \left( M_j' \right) = f \cdot \sum_j \left( g_j N'_j \right) = n_i, \text{ for } j, i = 1 \text{ to } I, \tag{25}
\]

or in terms of relative quantities more convenient for numerical work

\[
\frac{n_i}{n} = \sum_j g_j \left( \frac{N'_j}{N} \right). \tag{25a}
\]

Our empirical approach has involved the following steps:

- choosing a form for \( h \), and using it to determine \( N'_i \) defined in [24]
- using those in [25a] to iteratively determine the \( g_j \) values by StrCon which meet the sample allocation requirements \( n_i \) by sector, and hence the \( M_{ij} \) values from [24]
- computing the implied losses in efficiency due to weighting from [14], and their average over sectors, [15]
- comparing this average against (many) other choices of the function \( h \), and choosing the most efficient solution from among those computed.

The objective is to identify and choose the 'best' model, at least among those empirically evaluated. Comparing large numbers of numerical trials points at least to the direction we should be moving in the choice of the design parameters.

**Illustrations**

Here are examples of some of the forms which we have investigated and compared:

**Basic:** \( h=1 \), simply reducing [22] to its basic form (16):
\[
\frac{M_{x(j)}}{N_{x(j)}} = g_j.
\]

**Constant:** The size measure of units of the sector corresponding to the area's StrCon adjusted by a constant factor:

\[
\frac{M_{x(j)}}{N_{x(j)}} = g_j \left[ (1-x) + \left( 1 + c_j \right) x \right].
\]

**Linear:** The above size measures adjusted in linear proportion to x, the proportion of units in the area belonging to the sector corresponding to the area's StrCon:

\[
\frac{M_{x(j)}}{N_{x(j)}} = g_j \left[ (1-x) + \left( 1 + c_j \cdot x \right) x \right],
\]

**S-shaped:** The above with more elaborate variation with x; for instance:

\[
\frac{M_{x(j)}}{N_{x(j)}} = g_j \left[ (1-x) + \left( 1 + c_j \cdot x^2 \cdot (3-2x) \right) x \right],
\]

and so on. The last mentioned form appeared a good one at least for one survey tried.

The models above all assume \( h_j(x) \) to be of the form

\[
h_j(x) = c_j \cdot h(x),
\]

where function \( h(x) \) is independent of \( j \), i.e. is taken to be the same in all sectors in a given trial of the procedure. As for parameter \( c_j \) in the above, a simple choice we have tried is to take it as a function of the overall sampling rate required in the sector \((i=j)\) corresponding to StrCon \((j)\), such as:

\[
c_j = \left[ \left( \frac{f_i}{f_j} \right)^a - 1 \right], \quad \alpha \geq 0, \quad i = j.
\]

8. **Concluding remarks**

The specific examples given above must be taken as merely illustrative of the type of solutions we have looked into to the basic design problem in surveys of heterogeneous and unevenly distributed small-scale economic units. The solution depends on the nature of the population at hand, and we have used the approach sketched above in designing a number of samples, covering diverse types of surveys in different situations (countries). These have included surveys of agricultural and non-agricultural small-scale units, and even in a survey of schools where the objective was to control the sample allocation by ethnic group
but through a sample of schools on the basis of information on the schools' ethnic composition.

In this last example, it was not ethically permissible to identify and sample differentially students of different ethnic groups: all variations in the required overall sampling rates by ethnicity had to be achieved by appropriately adjusting the school selection probabilities (i.e. measures of size) according to pre-available information on the schools' ethnic composition.

In conclusion, one important and useful aspect of the above approach should be brought out. Meeting sample allocation requirements is a fundamental aspect of the design for surveys of the kind under discussion. Nevertheless, to the extent possible, the issue of sample allocation – involving the choice of the modification of measures of size \(M_k/N_k\) for instance – should be isolated from the structure and process of actual sample selection. Once assigned, the units always 'carry with themselves' their relative selection probabilities \(M_k/N_k\) irrespective of details of the selection process, and the required allocation is automatically ensured, at least in the statistically expected sense.

The advantage of such separation of allocation and selection aspects is that the structure and process of selection (stratification, multiple sampling stages, subsampling etc) can be determined flexibly, purely by considerations of sampling efficiency, and need not be constrained by the requirements of allocation.

When the two aspects overlap in practice – for instance when a whole design domain is to be over-sampled – that is a coincidental rather than an essential aspect of the design.
INFLUENCE OF NUMBERS OF GROUPED BALANCED HALF-SAMPLES ON EFFECTIVENESS OF VARIANCE ESTIMATION FOR COMPLEX SAMPLE SURVEYS

Jacek Jakubowski\textsuperscript{1}, Czesław Bracha\textsuperscript{2}

ABSTRACT

In this paper we consider the problem of approximative variance estimation in survey sampling. We investigate repeatedly grouped balanced half-sample method (Rao-Shao method). By means of simulation study, the complex sampling survey has been examined. Results confirm the usefulness of Rao-Shao method in practice of survey sampling.

\textit{Key words}: variance estimation; complex sample survey; repeatedly grouped balanced half-sample method; two-stage sample design; Hartley-Rao scheme.

1. Introduction

Survey sampling can be used in all cases in which we can not investigate the whole population, because it is impossible or the cost of investigation of all elements is too high. We are interested not only in estimations of parameters of the whole population or some of subpopulations, but also in precision of estimates, i.e. estimations of errors which we make. We express the precision of estimates in terms of variance or mean-square error of estimates of parameters which we are interested in\textsuperscript{3}. In literature on survey sampling, one can find formulas describing estimators of variance in the case when the probability sampling scheme (sampling design) and estimators are not simultaneously too complicated.

In practice the situation looks different. Different economic and administrative factors (cost, time and the like) force statisticians to construct the

\textsuperscript{1} Institute of Mathematics, Warsaw University and Research Centre for Economic and Statistical Studies of The Central Statistical Office and Polish Academy of Sciences.

\textsuperscript{2} Warsaw School of Economics and Research Centre for Economic and Statistical Studies of The Central Statistical Office and Polish Academy of Sciences.

\textsuperscript{3} Further, for simplicity, we do not distinguish between variance and mean-square error. Simple, we use term variance.
complicated sampling schemes. The sampling design as well as estimators are complex in many of important large-scale sample surveys realised in the Central Statistical Office (GUS) in Poland. The additional complication arise as a result of using random weights, which in turn is caused by poststratification or by nonresponse or by imperfection of the sampling frame. In these situations, it is impossible to use the well-known formulas. The attempt of obtaining formulas for variance estimators in complex surveys sampling leads to very clumsy transformations (usually connected with linearization methods using Taylor series expansions). Therefore calculations of the form of variance of estimator are very difficult or even impossible for the most of large and complex surveys carried on in GUS. The examples of such surveys are the household budget surveys and the labour force survey (BAEL). There are also situations, when either the sampling scheme is simple (simple proportional-to-size sampling without replacement) but appear zero second-order inclusion probabilities (for example Samiudin-Asad (1981) scheme) or the second-order inclusion probabilities are calculated with accuracy which decreases with increase of sampling fraction (for example Hartley-Rao (1962) scheme).

Statisticians, taking into account problems with variance estimation by classical methods, have been using simplifying variance estimation methods for a long time. P.C. Mahalanobis is a precursor of approximate variance estimates. He proposed (Mahalanobis (1939, 1944, 1946)) the methods of interpenetrating samples. Now, it is known as the method of independent random groups. It is a very simple method, but allows possibility of overlapping of random groups, which in the case of without replacement sampling schemes leads to decrease of efficiency the variance estimation of population parameters estimators.

Nowadays, besides independent random group method, there are different approximate variance estimation methods (see e.g. Wolter (1985), Särndal, Swensson and Wretman (1992)):

- dependent groups method,
- balanced repeated replication method (different name: balanced half-samples method),
- jack-knife method,
- bootstrap method,
- linearization method.

In this paper, we consider the balanced repeated replication (BRR) method. P.J. McCarthy has used this method for the first time (McCarthy (1966, 1969)) and then it has been developed by many authors (Kish and Frankel (1970), Wolter (1985), Rust (1985) etc.). The new approach to BRR methods was proposed by Rao and Shao (1996). They have suggested to estimate variance using the following methods. In the first step, the sample in each stratum is randomly divided into two groups, and than the BRR method is applied to the groups to calculate variance estimator. This procedure is independently repeated \( k \) times. In the second step, as an variance estimator, obtained by repeatedly grouped
balanced half-samples (RGBHS) method, is taken the average of the $k$ variance estimators obtained in the first step (one can see the analogy between RGBHS method and Norlén-Waller (1979) method). RGBHS method we call Rao-Shao method. An important question connected with application of Rao-Shao method is how fast variance of variance estimator of population parameter decreases, when we increase number of groups (i.e. number of replications). In particular we want to investigate the properties of efficiency of variance estimator – how efficiency improves with increasing of number of groups. It is impossible to solve this problem analytically. Only simulation studies allow us to understand the behaviour and some properties of the estimators. We know that results obtained by simulations study are not mathematical theorems with precise proofs. But simulation study with suitable planning experiment can provide some indication on usefulness of Rao-Shao method in practice.

We take two-stage sampling such that primary sampling units (PSUs) are stratified before sampling in order to close the simulation study to the real conditions of survey studies. In stratum PSUs are selected according to Hartley-Rao scheme with probabilities proportional-to-size of the second sampling units (SSUs) contain in PSU. At the second stage we draw the constant number of SSUs ($n_g=10$). Similar schemes are used in GUS in household budget surveys and in labour force survey, but with one difference – in both mentioned surveys rotation is added. In strata samples is self-weighting. Totals and ratios of totals over population and five subpopulation are estimated. Two kinds of estimators of parameters are used: with and without poststratification. Poststratification based on information on number of elements in subpopulations in strata. We consider the quite big population consisting of 100 000 elements, again to approximate situation to the real one.

2. Rao-Shao method

We start from the outline of Rao-Shao (RGBHS) method (see details in Rao and Shao (1996)). It is a generalisation of BRR method. BRR is a technique of approximative variance estimation for stratified sampling with two units drawn independently from each stratum. Using Hadamard matrices we form half-samples (they contain exactly one item from each stratum), which are in full orthogonal balance. If we estimate parameter $T$ using estimator $t$, then variance $V(t)$ we estimate by

$$v_{bs} = \frac{1}{m} \sum_{i=1}^{m} (t_i - t)^2,$$  

where $m$ is the number of half-samples, $t$ is the estimator of $T$ based on the whole sample and $t_i$ is the estimator of $T$ of the same form as $t$, based on $i$-th half-sample.
The next step is generalization of BRR method to samples which contains more than two elements in strata. To do this, the sample in each stratum is first randomly divided into two groups. The group is treated as one element and then BRR method is applied. But such procedure leads to incorrect conclusions. So the idea is to repeat independently grouping $D$ times, each time calculating the estimator $\nu_i^{(i)}$ of the form (1), and then as an variance estimator is taken

$$v = \frac{1}{D} \sum_{i=1}^{D} \nu_i^{(i)}.$$ (2)

Rao–Shao method preserve the simplicity of BRR method, since the same Hadamard matrix is used all the time.

3. Description of empirical study

We generate five populations, each contains 100 000 elements, as follows. For a given vector of mean $\bar{Y}$, standard deviation $S$ and correlation matrix $R$, we generate random numbers with six-dimensional gaussian distribution. The form of $\bar{Y}$, $S$, $R$ are given below:

$$\bar{Y}_1 = \begin{bmatrix} 0 \\ 0 \\ 10 \\ 10 \\ 20 \\ 30 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 3 \\ 10 \\ 10 \end{bmatrix}, \quad R_1 = \begin{bmatrix} 1 & 0.2 & 0.3 & 0.2 & 0.3 & 0.4 \\ 0.2 & 1 & 0.2 & 0.3 & 0.4 & 0.2 \\ 0.3 & 0.2 & 1 & 0.4 & 0.5 & 0.6 \\ 0.2 & 0.3 & 0.4 & 1 & 0.4 & 0.5 \\ 0.3 & 0.4 & 0.5 & 0.4 & 1 & 0.7 \\ 0.4 & 0.2 & 0.6 & 0.5 & 0.7 & 1 \end{bmatrix}.$$  

$$\bar{Y}_2 = \begin{bmatrix} 0 \\ 0 \\ 15 \\ 10 \\ 10 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 2 \\ 2 \\ 15 \\ 10 \\ 7.5 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 1 & 0.2 & 0.4 & 0.2 & 0.1 & 0.2 \\ 0.2 & 1 & 0.3 & 0.5 & 0.4 & 0.2 \\ 0.4 & 0.3 & 1 & 0.4 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.4 & 1 & 0.2 & 0.3 \\ 0.1 & 0.3 & 0.2 & 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 0.3 & 0.2 & 0.2 & 1 \end{bmatrix}.$$
Populations have the following parameters:

(a) populations I: $\overline{Y}_1, S_1, R_1$;
(b) populations II: $\overline{Y}_2, S_2, R_2$;
(c) populations III: $\overline{Y}_3, S_3, R_3$;
(d) populations IV: $\overline{Y}_4, S_4, R_4$;
(e) populations V: $\overline{Y}_5, S_5, R_5$.

In this way we obtain five matrices $Y = [Y_{ij}]_{i=1}^{100000}$ (one for each population). We have decided to close such five populations, since we want to investigate Rao-Shao method for different population. Populations differ in means, dispersion and correlation matrices.

Now we describe one population (rules are the same for all five populations). The first three variables of given six (first three columns of matrix $Y$) are only used for constitute strata ($Y_1$), primary sampling units ($Y_2$), and subpopulation ($Y_3$). The remaining variable $Y_4, Y_5, Y_6$ are variable of study. At first we divide population into subpopulations with help of $Y_3$ variable. The rows of matrix $Y$ we arrange in such a way that values of the third variable are nondecreasing order. Then, we take to succeeding subpopulation the following percentage of units study: 3%, 10%, 15%, 25%, 45% (i.e. the first 3% of items after ordering belongs to the first subpopulation, the second 10% to the second, etc.). So we have five disjoint subpopulation (values of variable $Y_3$ are replaced by numbers 1, 2, 3, 4, 5 hinted on number of subpopulation). Next the population was stratified into 6 strata using $Y_1$. The rows of modified matrix $Y$ are arranged in such a way that values of the first column constitute nondecreasing sequence. To the first stratum we take first $N_1 = 10000$ rows, to the second $N_2 = 12000$ rows, etc. Information on strata is given in Table 1. The following notation are used:

$N = 100000$ – the number of items in population (the number of SSUs),
$L = 6$ – the number of strata,
$M = 4000$ – the number of PSUs in the population,
The number of items in PSUs (\( l = 1, 2, \ldots, 6, h = 1, 2, \ldots, M_l \)) are rounded to the nearest integer number values of independent random variables with shifted exponential distribution with an additional restriction:

\[
15 \leq N_{lh} \leq 100.
\]

Taking minimal size of PSU equal to 15 guarantees that we can draw 10 SSUs at the second stage. It is common in GUS to combine small territorial statistical units in bigger cluster to assure the possibility of obtaining the constant size of SSUs drawing at the second stage (this gives self-weighting sample, if we use simple sampling proportional-to-size with or without replacement at the first stage. It is solutions which considerable simplifies methods of estimation and organization factor of survey). After fixing numbers \( N_{lh} \) we form PSUs. To do this, we use values of \( Y_l \). In each stratum we arrange the rules of the second column in descend order. For given \( l \), the first PSU consists of first \( N_{l1} \) rows of \( l \)-th stratum, the second PSU consists of next \( N_{l2} \) rows of \( l \)-th stratum, etc.
As we have mentioned, variable $Y_4, Y_5, Y_6$ are variables of study (so $Y_4$ is the first variable of study etc.). We denote by $Y_{lj}^{(i)}$, $i=1, 2, 3$ ($i$ – the number of variable of study), $l=1, \ldots, 6$, $h=1, \ldots, M_i$, $j=1, \ldots, N_{lh}$, values of variables of study. To avoid negative values (economic variables are usually positive) we take the absolute value of $Y_{lj}^{(i)}$. Therefore, variables of study have distributions with right skewness. It is consistent with observation that variables of economic character have generally such type of distribution. So by $X_{lj}^{(i+3k)}$ we denote value of $i$-th variable in $k$-th subpopulation ($k=0$ corresponds to the whole population) in $j$-th SSU belonging to $h$-th PSU in $l$-th stratum (for example, the superscript 17 means that we investigate second variable in the fifth subpopulation). The sizes of subpopulations and strata are given in Table 2.

**Table 2.**

<table>
<thead>
<tr>
<th>$l$</th>
<th>$k=0$</th>
<th>$k=1$</th>
<th>$k=2$</th>
<th>$k=3$</th>
<th>$k=4$</th>
<th>$k=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100 000</td>
<td>2 926</td>
<td>10 242</td>
<td>15 022</td>
<td>24 881</td>
<td>46 929</td>
</tr>
<tr>
<td>1</td>
<td>10 000</td>
<td>288</td>
<td>1 026</td>
<td>1 504</td>
<td>2 486</td>
<td>4 696</td>
</tr>
<tr>
<td>2</td>
<td>12 000</td>
<td>334</td>
<td>1 235</td>
<td>1 817</td>
<td>2 977</td>
<td>5 637</td>
</tr>
<tr>
<td>3</td>
<td>15 000</td>
<td>449</td>
<td>1 525</td>
<td>2 264</td>
<td>3 728</td>
<td>7 034</td>
</tr>
<tr>
<td>4</td>
<td>21 000</td>
<td>702</td>
<td>2 130</td>
<td>3 149</td>
<td>5 237</td>
<td>9 782</td>
</tr>
<tr>
<td>5</td>
<td>18 000</td>
<td>471</td>
<td>1 863</td>
<td>2 701</td>
<td>4 481</td>
<td>8 484</td>
</tr>
<tr>
<td>6</td>
<td>24 000</td>
<td>682</td>
<td>2 463</td>
<td>3 587</td>
<td>5 972</td>
<td>11 296</td>
</tr>
</tbody>
</table>

We use information contained in this table to poststratification.

The Hartley-Rao scheme was used at the first stage (in strata) with inclusion probabilities proportional to $N_{lj}$. At the second stage simple sampling without replacement with constant size of SSUs, $n_0=10$ (recall that this gives self-weighting sample in strata) have been used.

We are interested in totals over population and subpopulations:

$$X^{(i+3k)} = \sum_{l=1}^{6} \sum_{h=1}^{M_i} X_{lj}^{(i+3k)} , \quad (i=1,2,3, k=0,1,...,5)$$

where

$$X_{lj}^{(i+3k)} = \sum_{j=1}^{N_{lj}} X_{lj}^{(i+3k)} ,$$

and ratio of them
\[ R^{(i+3k,l'+3k)} = \frac{X^{(i+3k)}}{X^{(l'+3k)}}, \quad (i,l'=1,2,3, i \neq l', k=0,1,\ldots,5). \]

Denote by \[ x^{(i+3k)}_{(l'g)} \quad (i=1,2,3, k=0,1,\ldots,5, l=1,\ldots,6, g=1,\ldots,m, f=1,\ldots,10) \] value of \( i \)-th variable for unit from \( f \)-th SSu in \( g \)-th PSU in \( l \)-th stratum and \( k \)-th subpopulation.

Unbiased estimator of \[ X^{(i+3k)} \] is of the form
\[ \hat{X}^{(i+3k)} = \sum_{l=1}^{6} \frac{N_l}{10m_l} \sum_{g=1}^{m_l} \sum_{f=1}^{10} x^{(i+3k)}_{(l'g)} \quad (i=1,2,3, k=0,1,\ldots,5), \]
with variance given by approximative formula \( (i=1,2,\ldots,18) \)
\[ V(\hat{X}^{(i)}) \approx \sum_{l=1}^{6} \left[ \sum_{h=1}^{M_l} \pi_{lh} \left( 1 - \frac{m_l - 1}{m_l} \frac{\pi_{lh}}{\bar{\pi}_{lh}} \left( \frac{X_{[h]}^{(i)}}{\bar{\pi}_{lh}} - X_{[h]}^{(i)} \right)^2 \right) + \sum_{h=1}^{M_l} N_{lh} (N_{lh} - 10) \frac{S_{2h}^{(i)}}{10\pi_{lh}} \right], \quad (3) \]
where \( \pi_{lh} = m_l N_{lh} / N_l \) is an inclusion probability of \( h \)-th PSU in \( l \)-th stratum and
\[ S_{2h}^{(i)} = \frac{1}{N_{lh} - 1} \sum_{j=1}^{N_{lh}} (X_{[h]}^{(i)} - \bar{X}_{[h]}^{(i)})^2, \]
\[ \bar{X}_{[h]}^{(i)} = \frac{1}{N_{lh}} \sum_{i=1}^{N_{lh}} X_{[h]}^{(i)}, \]
\[ X_{[h]}^{(i)} = \sum_{h=1}^{M_l} X_{[h]}^{(i)}. \]

Estimator of \[ R^{(i+3k,l'+3k)} \] is given by
\[ \hat{R}^{(i+3k,l'+3k)} = \frac{\hat{X}^{(i+3k)}}{\hat{X}^{(l'+3k)}}. \]
Moreover
\[ V(\hat{R}^{(i+3k,l'+3k)}) = \sum_{l=1}^{6} V(\hat{R}^{(i+3k,l'+3k)}), \quad (4) \]
where
\[ 1 \] Hence, \( \pi_{lh} = \pi_{lh} \cdot 10 / N_{lh} = 10m_l / N_l \). Therefore the sample is self-weighting in the strata.
\[
V\left(\hat{R}_{j}^{(i+3k,j'+3k)}\right) \approx V\left(\hat{X}_{j}^{(i+3k)}\right) - 2R_{j}^{(i+3k,j'+3k)} \text{Cov}\left(\hat{X}_{j}^{(i+3k)} , \hat{X}_{j'}^{(j'+3k)}\right) + \left(\text{Cov}\left(\hat{X}_{j}^{(i+3k)}\right)\right)^{2} V\left(\hat{X}_{j'}^{(j'+3k)}\right),
\]

\[
\text{Cov}\left(\hat{X}_{j}^{(i+3k)} , \hat{X}_{j'}^{(j'+3k)}\right) \approx \sum_{h=1}^{M} \pi_{ih} \left\{1 - \frac{m_{i} - 1}{m_{j}} \pi_{ih} \left(\frac{X_{ih}^{(i+3k)}}{\pi_{ih}} - \frac{X_{ih}^{(i+3k)}}{m_{j}} \right) \right\} + \sum_{h=1}^{M} N_{ih} \left(\pi_{ih} - 10\right) \frac{S_{j}^{(i+3k,j'+3k)}}{10 \pi_{ih}},
\]

and

\[
S_{j}^{(i+3k,j'+3k)} = \frac{1}{N_{ih} - 1} \sum_{j=1}^{N_{ih}} \left(X_{ihj}^{(i+3k)} - \bar{X}_{ih}^{(i+3k)}\right) \left(X_{ihj}^{(j'+3k)} - \bar{X}_{ih}^{(j'+3k)}\right).
\]

The relative standard error of estimators of parameters mentioned above have been calculated according to the formula

\[
\delta = \frac{2\sqrt{V(T)}}{T},
\]

where \(T\) is an estimated parameter, \(t\) – his estimator. The relative standard errors are given in Tables 3 and 4.

**Table 3.** The relative standard error of estimators of total \(X^{(i+3k)}\) in populations and subpopulations

<table>
<thead>
<tr>
<th>(a)</th>
<th>(k=0)</th>
<th>(i=1)</th>
<th>(i=2)</th>
<th>(i=3)</th>
<th>(k=1)</th>
<th>(i=1)</th>
<th>(i=2)</th>
<th>(i=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0,01</td>
<td>0,02</td>
<td>0,01</td>
<td>1</td>
<td>0,14</td>
<td>0,15</td>
<td>0,14</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0,03</td>
<td>0,03</td>
<td>0,02</td>
<td>2</td>
<td>0,22</td>
<td>0,22</td>
<td>0,21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>3</td>
<td>0,17</td>
<td>0,17</td>
<td>0,18</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0,03</td>
<td>0,02</td>
<td>0,02</td>
<td>4</td>
<td>0,22</td>
<td>0,22</td>
<td>0,21</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0,01</td>
<td>0,01</td>
<td></td>
<td>5</td>
<td>0,17</td>
<td>0,18</td>
<td>0,18</td>
<td></td>
</tr>
</tbody>
</table>

\(a\) - it is the number of population
### Table 4. The relative standard error of estimators of ratio of totals $R^{(i+3k, i'+3k)}$ in populations and subpopulations

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=2$</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0,10</td>
<td>0,11</td>
<td>0,10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0,11</td>
<td>0,11</td>
<td>0,10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0,08</td>
<td>0,08</td>
<td>0,09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0,11</td>
<td>0,10</td>
<td>0,10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0,08</td>
<td>0,08</td>
<td>0,08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=3$</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0,07</td>
<td>0,08</td>
<td>0,07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0,09</td>
<td>0,09</td>
<td>0,08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0,07</td>
<td>0,07</td>
<td>0,07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0,09</td>
<td>0,08</td>
<td>0,08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0,07</td>
<td>0,07</td>
<td>0,07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=4$</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0,06</td>
<td>0,07</td>
<td>0,06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0,07</td>
<td>0,06</td>
<td>0,06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0,05</td>
<td>0,05</td>
<td>0,05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0,07</td>
<td>0,06</td>
<td>0,06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0,05</td>
<td>0,05</td>
<td>0,05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=5$</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0,04</td>
<td>0,05</td>
<td>0,04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0,05</td>
<td>0,04</td>
<td>0,04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0,03</td>
<td>0,03</td>
<td>0,03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0,05</td>
<td>0,04</td>
<td>0,04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0,03</td>
<td>0,03</td>
<td>0,03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=0$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1,2)$</td>
<td>$(1,3)$</td>
<td>$(2,1)$</td>
<td>$(2,3)$</td>
<td>$(3,1)$</td>
<td>$(3,2)$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0,02</td>
<td>0,01</td>
<td>0,02</td>
<td>0,02</td>
<td>0,01</td>
<td>0,02</td>
</tr>
<tr>
<td>2</td>
<td>0,03</td>
<td>0,03</td>
<td>0,03</td>
<td>0,03</td>
<td>0,03</td>
<td>0,03</td>
</tr>
<tr>
<td>3</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
</tr>
<tr>
<td>4</td>
<td>0,02</td>
<td>0,03</td>
<td>0,02</td>
<td>0,02</td>
<td>0,03</td>
<td>0,02</td>
</tr>
<tr>
<td>5</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=1$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$(1,2)$</td>
<td>$(1,3)$</td>
<td>$(2,1)$</td>
<td>$(2,3)$</td>
<td>$(3,1)$</td>
<td>$(3,2)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1,2)</td>
<td>(1,3)</td>
<td>(2,1)</td>
<td>(2,3)</td>
<td>(3,1)</td>
<td>(3,2)</td>
</tr>
<tr>
<td>---</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>0.18</td>
<td>0.19</td>
<td>0.17</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.08</td>
<td>0.06</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>0.12</td>
<td>0.14</td>
<td>0.12</td>
<td>0.13</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
<td>0.05</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=2$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.04</td>
<td>0.06</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.09</td>
<td>0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.06</td>
<td>0.07</td>
<td>0.06</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=3$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=4$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$a$</th>
<th>$k=5$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>
The smallest errors are for the whole population and change between 1% and 3%. In subpopulations these errors are significantly higher, but decrease as the size of subpopulation increase (the biggest for $k=1$, the smallest for $k=5$; it is not surprising, since as we see from Table 2, the first subpopulation is the smallest one, the five subpopulation is the biggest one). In strata errors are considerable higher, but they are acceptable taking into account the size of sampling in strata.

Our main goal has been the investigation of dependence of efficiency of variance estimators of the form (3) and (4) on the number of repetition (i.e. the number of groups, which we constitute). Denote by $V_p^{(t)}$ estimation of variance $V(t)$ obtaining in $p$-th iteration ($p=1,...,2000$). At first we calculate the average

$$
\bar{V}^{(t)} = \frac{1}{2000} \sum_{p=1}^{2000} V_p^{(t)},
$$

which is an approximation of expected value of estimator used to estimate of $V(t)$.

Next, we derive

$$
S_t^2 = \frac{1}{1999} \sum_{p=1}^{2000} (V_p^{(t)} - \bar{V}^{(t)})^2.
$$

It estimated the true variance of estimator $\hat{V}(t)$ to the degree of precision obtained with 2000 iterations. Then we calculate the approximation of coefficient of variation according to formula

$$
\nu(\hat{V}(t)) = \frac{\sqrt{S_t^2}}{\bar{V}(t)}.
$$

These coefficients are recognized as a measure of efficiency of estimator $t$. The less the coefficient $\nu(\hat{V}(t))$ is, the more effective is the estimator $\hat{V}(t)$. Results of the simulations are given in tables 5–10. We start from description of results for estimation of totals (tables 5–9).

If $D=1$ (the case of grouped BRR), then relative efficiency are for all populations fairly similar, although populations are different. The biggest values are for IV-th population. For this population all quantities obtained in simulation study are slightly higher than corresponding values in other populations.

We can see, as a general rule, that if we increase the number of repetition $D$, then the relative efficiency of estimators have been improved in all cases. For $D=5$ relative efficiency of estimator of totals decreases twice, for $D=10$ their

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>0,03</th>
<th>0,03</th>
<th>0,03</th>
<th>0,02</th>
<th>0,03</th>
<th>0,02</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
<td>0,01</td>
</tr>
</tbody>
</table>
values are equal to 30% of values for $D=1$. For $D=100$ repetitions relative efficiency of estimators of totals without poststratification are less than 0.15 in all cases besides quantities for the first subpopulation (it is the least subpopulation and contains 3% of all items) in populations II-nd and IV-th. If we consider estimators of totals with poststratification, then relative efficiency is always slightly higher than relative efficiency of estimators of totals without poststratification, but they are again less than 0.15 apart from the first subpopulation.

Table 5. Relative efficiency of estimators of totals without and with poststratification, over first population and its subpopulations

<table>
<thead>
<tr>
<th>$D$</th>
<th>$k=0$</th>
<th></th>
<th>$k=1$</th>
<th></th>
<th>$k=2$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
</tr>
<tr>
<td>1</td>
<td>0.701</td>
<td>0.732</td>
<td>0.688</td>
<td>0.669</td>
<td>0.731</td>
<td>0.712</td>
</tr>
<tr>
<td>5</td>
<td>0.334</td>
<td>0.335</td>
<td>0.307</td>
<td>0.321</td>
<td>0.345</td>
<td>0.337</td>
</tr>
<tr>
<td>10</td>
<td>0.244</td>
<td>0.245</td>
<td>0.231</td>
<td>0.241</td>
<td>0.263</td>
<td>0.251</td>
</tr>
<tr>
<td>15</td>
<td>0.206</td>
<td>0.211</td>
<td>0.196</td>
<td>0.209</td>
<td>0.229</td>
<td>0.213</td>
</tr>
<tr>
<td>20</td>
<td>0.188</td>
<td>0.188</td>
<td>0.172</td>
<td>0.189</td>
<td>0.210</td>
<td>0.194</td>
</tr>
<tr>
<td>50</td>
<td>0.139</td>
<td>0.141</td>
<td>0.130</td>
<td>0.142</td>
<td>0.167</td>
<td>0.148</td>
</tr>
<tr>
<td>100</td>
<td>0.116</td>
<td>0.118</td>
<td>0.111</td>
<td>0.127</td>
<td>0.156</td>
<td>0.133</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D$</th>
<th>$k=3$</th>
<th></th>
<th>$k=4$</th>
<th></th>
<th>$k=5$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
</tr>
<tr>
<td>1</td>
<td>0.673</td>
<td>0.717</td>
<td>0.728</td>
<td>0.678</td>
<td>0.702</td>
<td>0.684</td>
</tr>
<tr>
<td>5</td>
<td>0.303</td>
<td>0.321</td>
<td>0.320</td>
<td>0.314</td>
<td>0.315</td>
<td>0.319</td>
</tr>
<tr>
<td>10</td>
<td>0.225</td>
<td>0.234</td>
<td>0.235</td>
<td>0.231</td>
<td>0.236</td>
<td>0.234</td>
</tr>
<tr>
<td>15</td>
<td>0.188</td>
<td>0.200</td>
<td>0.196</td>
<td>0.195</td>
<td>0.198</td>
<td>0.198</td>
</tr>
<tr>
<td>20</td>
<td>0.169</td>
<td>0.179</td>
<td>0.176</td>
<td>0.172</td>
<td>0.178</td>
<td>0.176</td>
</tr>
<tr>
<td>50</td>
<td>0.128</td>
<td>0.138</td>
<td>0.133</td>
<td>0.128</td>
<td>0.135</td>
<td>0.131</td>
</tr>
</tbody>
</table>
Table 6. Relative efficiency of estimators of totals without and with poststratification, over second population and its subpopulations

<table>
<thead>
<tr>
<th>D</th>
<th>k=0</th>
<th>k=1</th>
<th>k=2</th>
<th>k=3</th>
<th>k=4</th>
<th>k=5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i=1</td>
<td>i=2</td>
<td>i=3</td>
<td>i=1</td>
<td>i=2</td>
<td>i=3</td>
</tr>
<tr>
<td>1</td>
<td>0,741</td>
<td>0,693</td>
<td>0,720</td>
<td>0,710</td>
<td>0,712</td>
<td>0,706</td>
</tr>
<tr>
<td></td>
<td>0,740</td>
<td>0,687</td>
<td>0,712</td>
<td>0,710</td>
<td>0,710</td>
<td>0,706</td>
</tr>
<tr>
<td>5</td>
<td>0,349</td>
<td>0,323</td>
<td>0,326</td>
<td>0,360</td>
<td>0,334</td>
<td>0,346</td>
</tr>
<tr>
<td></td>
<td>0,348</td>
<td>0,320</td>
<td>0,324</td>
<td>0,360</td>
<td>0,334</td>
<td>0,346</td>
</tr>
<tr>
<td>10</td>
<td>0,261</td>
<td>0,240</td>
<td>0,243</td>
<td>0,282</td>
<td>0,266</td>
<td>0,263</td>
</tr>
<tr>
<td></td>
<td>0,259</td>
<td>0,237</td>
<td>0,241</td>
<td>0,282</td>
<td>0,266</td>
<td>0,263</td>
</tr>
<tr>
<td>15</td>
<td>0,219</td>
<td>0,203</td>
<td>0,210</td>
<td>0,250</td>
<td>0,235</td>
<td>0,229</td>
</tr>
<tr>
<td></td>
<td>0,218</td>
<td>0,201</td>
<td>0,208</td>
<td>0,250</td>
<td>0,235</td>
<td>0,229</td>
</tr>
<tr>
<td>20</td>
<td>0,197</td>
<td>0,180</td>
<td>0,186</td>
<td>0,230</td>
<td>0,218</td>
<td>0,217</td>
</tr>
<tr>
<td></td>
<td>0,196</td>
<td>0,179</td>
<td>0,184</td>
<td>0,230</td>
<td>0,218</td>
<td>0,217</td>
</tr>
<tr>
<td>50</td>
<td>0,144</td>
<td>0,135</td>
<td>0,133</td>
<td>0,193</td>
<td>0,183</td>
<td>0,180</td>
</tr>
<tr>
<td></td>
<td>0,144</td>
<td>0,133</td>
<td>0,132</td>
<td>0,193</td>
<td>0,183</td>
<td>0,180</td>
</tr>
<tr>
<td>100</td>
<td>0,123</td>
<td>0,115</td>
<td>0,113</td>
<td>0,184</td>
<td>0,174</td>
<td>0,168</td>
</tr>
<tr>
<td></td>
<td>0,123</td>
<td>0,114</td>
<td>0,112</td>
<td>0,184</td>
<td>0,174</td>
<td>0,168</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D</th>
<th>k=3</th>
<th>k=4</th>
<th>k=5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i=1</td>
<td>i=2</td>
<td>i=3</td>
</tr>
<tr>
<td>1</td>
<td>0,701</td>
<td>0,669</td>
<td>0,713</td>
</tr>
<tr>
<td></td>
<td>0,699</td>
<td>0,680</td>
<td>0,720</td>
</tr>
<tr>
<td>5</td>
<td>0,324</td>
<td>0,313</td>
<td>0,311</td>
</tr>
<tr>
<td></td>
<td>0,320</td>
<td>0,308</td>
<td>0,308</td>
</tr>
<tr>
<td>10</td>
<td>0,244</td>
<td>0,231</td>
<td>0,236</td>
</tr>
<tr>
<td></td>
<td>0,241</td>
<td>0,226</td>
<td>0,232</td>
</tr>
<tr>
<td>15</td>
<td>0,212</td>
<td>0,195</td>
<td>0,196</td>
</tr>
<tr>
<td></td>
<td>0,208</td>
<td>0,195</td>
<td>0,196</td>
</tr>
<tr>
<td>20</td>
<td>0,194</td>
<td>0,182</td>
<td>0,185</td>
</tr>
<tr>
<td></td>
<td>0,189</td>
<td>0,178</td>
<td>0,177</td>
</tr>
<tr>
<td>50</td>
<td>0,157</td>
<td>0,144</td>
<td>0,145</td>
</tr>
<tr>
<td></td>
<td>0,153</td>
<td>0,140</td>
<td>0,139</td>
</tr>
</tbody>
</table>
Table 7. Relative efficiency of estimators of totals without and with poststratification, over third population and its subpopulations

<table>
<thead>
<tr>
<th>$D$</th>
<th>$k=0$</th>
<th>$k=1$</th>
<th>$k=2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
</tr>
<tr>
<td>1</td>
<td>0,742</td>
<td>0,711</td>
<td>0,717</td>
</tr>
<tr>
<td></td>
<td>0,758</td>
<td>0,730</td>
<td>0,739</td>
</tr>
<tr>
<td>5</td>
<td>0,357</td>
<td>0,344</td>
<td>0,325</td>
</tr>
<tr>
<td></td>
<td>0,357</td>
<td>0,344</td>
<td>0,325</td>
</tr>
<tr>
<td>10</td>
<td>0,260</td>
<td>0,245</td>
<td>0,238</td>
</tr>
<tr>
<td></td>
<td>0,265</td>
<td>0,254</td>
<td>0,244</td>
</tr>
<tr>
<td>15</td>
<td>0,221</td>
<td>0,205</td>
<td>0,205</td>
</tr>
<tr>
<td></td>
<td>0,225</td>
<td>0,213</td>
<td>0,211</td>
</tr>
<tr>
<td>20</td>
<td>0,199</td>
<td>0,181</td>
<td>0,183</td>
</tr>
<tr>
<td></td>
<td>0,203</td>
<td>0,187</td>
<td>0,189</td>
</tr>
<tr>
<td>50</td>
<td>0,147</td>
<td>0,142</td>
<td>0,138</td>
</tr>
<tr>
<td></td>
<td>0,147</td>
<td>0,142</td>
<td>0,138</td>
</tr>
<tr>
<td>100</td>
<td>0,120</td>
<td>0,115</td>
<td>0,114</td>
</tr>
<tr>
<td></td>
<td>0,123</td>
<td>0,119</td>
<td>0,117</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$D$</th>
<th>$k=3$</th>
<th>$k=4$</th>
<th>$k=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
</tr>
<tr>
<td>1</td>
<td>0,661</td>
<td>0,681</td>
<td>0,700</td>
</tr>
<tr>
<td></td>
<td>0,668</td>
<td>0,688</td>
<td>0,709</td>
</tr>
<tr>
<td>5</td>
<td>0,296</td>
<td>0,299</td>
<td>0,301</td>
</tr>
<tr>
<td></td>
<td>0,303</td>
<td>0,307</td>
<td>0,307</td>
</tr>
<tr>
<td>10</td>
<td>0,219</td>
<td>0,219</td>
<td>0,224</td>
</tr>
<tr>
<td></td>
<td>0,227</td>
<td>0,228</td>
<td>0,230</td>
</tr>
<tr>
<td>15</td>
<td>0,185</td>
<td>0,185</td>
<td>0,187</td>
</tr>
<tr>
<td></td>
<td>0,195</td>
<td>0,196</td>
<td>0,197</td>
</tr>
<tr>
<td>20</td>
<td>0,167</td>
<td>0,165</td>
<td>0,168</td>
</tr>
<tr>
<td></td>
<td>0,178</td>
<td>0,175</td>
<td>0,179</td>
</tr>
<tr>
<td>50</td>
<td>0,129</td>
<td>0,126</td>
<td>0,129</td>
</tr>
<tr>
<td></td>
<td>0,137</td>
<td>0,134</td>
<td>0,138</td>
</tr>
<tr>
<td>100</td>
<td>0,110</td>
<td>0,106</td>
<td>0,109</td>
</tr>
</tbody>
</table>
Table 8. Relative efficiency of estimators of totals without and with poststratification, over fourth population and its subpopulations

<table>
<thead>
<tr>
<th>$D$</th>
<th>$k=0$</th>
<th>$k=1$</th>
<th>$k=2$</th>
<th>$k=3$</th>
<th>$k=4$</th>
<th>$k=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
<td>$i=1$</td>
<td>$i=2$</td>
<td>$i=3$</td>
</tr>
<tr>
<td>1</td>
<td>0.857</td>
<td>0.798</td>
<td>0.722</td>
<td>0.738</td>
<td>0.820</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td>0.859</td>
<td>0.806</td>
<td>0.732</td>
<td>0.859</td>
<td>0.928</td>
<td>0.844</td>
</tr>
<tr>
<td>5</td>
<td>0.401</td>
<td>0.364</td>
<td>0.333</td>
<td>0.374</td>
<td>0.389</td>
<td>0.376</td>
</tr>
<tr>
<td></td>
<td>0.402</td>
<td>0.369</td>
<td>0.338</td>
<td>0.430</td>
<td>0.450</td>
<td>0.423</td>
</tr>
<tr>
<td>10</td>
<td>0.293</td>
<td>0.268</td>
<td>0.250</td>
<td>0.291</td>
<td>0.296</td>
<td>0.284</td>
</tr>
<tr>
<td></td>
<td>0.295</td>
<td>0.273</td>
<td>0.254</td>
<td>0.338</td>
<td>0.350</td>
<td>0.336</td>
</tr>
<tr>
<td>15</td>
<td>0.244</td>
<td>0.227</td>
<td>0.212</td>
<td>0.256</td>
<td>0.257</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>0.246</td>
<td>0.230</td>
<td>0.214</td>
<td>0.301</td>
<td>0.313</td>
<td>0.297</td>
</tr>
<tr>
<td>20</td>
<td>0.217</td>
<td>0.202</td>
<td>0.185</td>
<td>0.234</td>
<td>0.235</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>0.219</td>
<td>0.205</td>
<td>0.187</td>
<td>0.281</td>
<td>0.294</td>
<td>0.281</td>
</tr>
<tr>
<td>50</td>
<td>0.155</td>
<td>0.152</td>
<td>0.137</td>
<td>0.191</td>
<td>0.183</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>0.157</td>
<td>0.154</td>
<td>0.139</td>
<td>0.239</td>
<td>0.245</td>
<td>0.241</td>
</tr>
<tr>
<td>100</td>
<td>0.129</td>
<td>0.125</td>
<td>0.116</td>
<td>0.181</td>
<td>0.170</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td>0.130</td>
<td>0.127</td>
<td>0.118</td>
<td>0.228</td>
<td>0.231</td>
<td>0.224</td>
</tr>
</tbody>
</table>
Table 9. Relative efficiency of estimators of totals without and with poststratification, over fifth population and its subpopulations

<table>
<thead>
<tr>
<th></th>
<th>(k=0)</th>
<th>(i=1)</th>
<th>(i=2)</th>
<th>(i=3)</th>
<th>(k=1)</th>
<th>(i=1)</th>
<th>(i=2)</th>
<th>(i=3)</th>
<th>(k=2)</th>
<th>(i=1)</th>
<th>(i=2)</th>
<th>(i=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(D)</td>
<td>0,741</td>
<td>0,707</td>
<td>0,695</td>
<td>0,695</td>
<td>0,728</td>
<td>0,719</td>
<td>0,643</td>
<td>0,666</td>
<td>0,663</td>
<td>0,728</td>
<td>0,719</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,762</td>
<td>0,753</td>
<td>0,726</td>
<td>0,788</td>
<td>0,830</td>
<td>0,806</td>
<td>0,666</td>
<td>0,687</td>
<td>0,689</td>
<td>0,695</td>
<td>0,695</td>
</tr>
<tr>
<td>5</td>
<td>(k=3)</td>
<td>0,352</td>
<td>0,327</td>
<td>0,314</td>
<td>0,330</td>
<td>0,340</td>
<td>0,342</td>
<td>0,309</td>
<td>0,319</td>
<td>0,317</td>
<td>0,378</td>
<td>0,388</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,358</td>
<td>0,348</td>
<td>0,326</td>
<td>0,378</td>
<td>0,393</td>
<td>0,389</td>
<td>0,325</td>
<td>0,336</td>
<td>0,335</td>
<td>0,340</td>
<td>0,342</td>
</tr>
<tr>
<td>10</td>
<td>(k=4)</td>
<td>0,260</td>
<td>0,239</td>
<td>0,237</td>
<td>0,244</td>
<td>0,250</td>
<td>0,251</td>
<td>0,225</td>
<td>0,234</td>
<td>0,230</td>
<td>0,288</td>
<td>0,300</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,266</td>
<td>0,258</td>
<td>0,247</td>
<td>0,289</td>
<td>0,300</td>
<td>0,300</td>
<td>0,243</td>
<td>0,253</td>
<td>0,249</td>
<td>0,290</td>
<td>0,300</td>
</tr>
<tr>
<td>15</td>
<td>(k=5)</td>
<td>0,221</td>
<td>0,203</td>
<td>0,201</td>
<td>0,209</td>
<td>0,211</td>
<td>0,213</td>
<td>0,193</td>
<td>0,200</td>
<td>0,196</td>
<td>0,262</td>
<td>0,266</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,227</td>
<td>0,219</td>
<td>0,209</td>
<td>0,252</td>
<td>0,260</td>
<td>0,260</td>
<td>0,213</td>
<td>0,221</td>
<td>0,215</td>
<td>0,290</td>
<td>0,300</td>
</tr>
<tr>
<td>20</td>
<td>(k=6)</td>
<td>0,200</td>
<td>0,184</td>
<td>0,178</td>
<td>0,188</td>
<td>0,190</td>
<td>0,194</td>
<td>0,174</td>
<td>0,177</td>
<td>0,175</td>
<td>0,242</td>
<td>0,246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,205</td>
<td>0,198</td>
<td>0,186</td>
<td>0,232</td>
<td>0,240</td>
<td>0,243</td>
<td>0,193</td>
<td>0,198</td>
<td>0,194</td>
<td>0,245</td>
<td>0,249</td>
</tr>
<tr>
<td>50</td>
<td>(k=7)</td>
<td>0,145</td>
<td>0,137</td>
<td>0,132</td>
<td>0,139</td>
<td>0,139</td>
<td>0,145</td>
<td>0,126</td>
<td>0,127</td>
<td>0,127</td>
<td>0,315</td>
<td>0,318</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,149</td>
<td>0,149</td>
<td>0,138</td>
<td>0,187</td>
<td>0,195</td>
<td>0,201</td>
<td>0,148</td>
<td>0,152</td>
<td>0,149</td>
<td>0,319</td>
<td>0,322</td>
</tr>
<tr>
<td>100</td>
<td>(k=8)</td>
<td>0,120</td>
<td>0,115</td>
<td>0,111</td>
<td>0,122</td>
<td>0,121</td>
<td>0,129</td>
<td>0,105</td>
<td>0,105</td>
<td>0,106</td>
<td>0,410</td>
<td>0,413</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0,124</td>
<td>0,125</td>
<td>0,116</td>
<td>0,172</td>
<td>0,179</td>
<td>0,185</td>
<td>0,129</td>
<td>0,132</td>
<td>0,130</td>
<td>0,416</td>
<td>0,419</td>
</tr>
</tbody>
</table>

**Notes:**

- The table presents the relative efficiency of estimators of totals for different levels of poststratification and subpopulation sizes.
- The entries represent the efficiency ratios for various values of \(D\) and \(k\).
For relative efficiency of estimators of ratios of totals (see table 10) we can infer analogous corollaries as described above for relative efficiency of estimators of total. Besides, for $D=100$ results for estimation with poststratification are nearly the same as for estimation without poststratification. For the sake of place we place only results for the whole population ($k=0$).

Table 10. Relative efficiency of estimators of ratios of totals without and with poststratification

### Population I

<table>
<thead>
<tr>
<th>$D$</th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>$i_6$</th>
<th>$i_7$</th>
<th>$i_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0,741</td>
<td>0,742</td>
<td>0,666</td>
<td>0,666</td>
<td>0,716</td>
<td>0,716</td>
<td>0,745</td>
<td>0,746</td>
</tr>
<tr>
<td>5</td>
<td>0,326</td>
<td>0,327</td>
<td>0,316</td>
<td>0,317</td>
<td>0,315</td>
<td>0,316</td>
<td>0,344</td>
<td>0,345</td>
</tr>
<tr>
<td>10</td>
<td>0,242</td>
<td>0,243</td>
<td>0,236</td>
<td>0,236</td>
<td>0,233</td>
<td>0,234</td>
<td>0,254</td>
<td>0,254</td>
</tr>
<tr>
<td>15</td>
<td>0,207</td>
<td>0,208</td>
<td>0,195</td>
<td>0,195</td>
<td>0,198</td>
<td>0,199</td>
<td>0,211</td>
<td>0,212</td>
</tr>
<tr>
<td>20</td>
<td>0,185</td>
<td>0,186</td>
<td>0,177</td>
<td>0,177</td>
<td>0,177</td>
<td>0,177</td>
<td>0,189</td>
<td>0,190</td>
</tr>
<tr>
<td>50</td>
<td>0,140</td>
<td>0,141</td>
<td>0,131</td>
<td>0,131</td>
<td>0,132</td>
<td>0,133</td>
<td>0,141</td>
<td>0,142</td>
</tr>
<tr>
<td>100</td>
<td>0,120</td>
<td>0,121</td>
<td>0,110</td>
<td>0,111</td>
<td>0,112</td>
<td>0,113</td>
<td>0,117</td>
<td>0,118</td>
</tr>
</tbody>
</table>

### Population II

<table>
<thead>
<tr>
<th>$D$</th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>$i_6$</th>
<th>$i_7$</th>
<th>$i_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0,683</td>
<td>0,687</td>
<td>0,739</td>
<td>0,739</td>
<td>0,679</td>
<td>0,684</td>
<td>0,696</td>
<td>0,697</td>
</tr>
<tr>
<td>5</td>
<td>0,323</td>
<td>0,324</td>
<td>0,336</td>
<td>0,336</td>
<td>0,320</td>
<td>0,321</td>
<td>0,317</td>
<td>0,319</td>
</tr>
<tr>
<td>10</td>
<td>0,236</td>
<td>0,237</td>
<td>0,246</td>
<td>0,247</td>
<td>0,233</td>
<td>0,234</td>
<td>0,232</td>
<td>0,234</td>
</tr>
<tr>
<td>15</td>
<td>0,199</td>
<td>0,200</td>
<td>0,209</td>
<td>0,210</td>
<td>0,197</td>
<td>0,198</td>
<td>0,198</td>
<td>0,199</td>
</tr>
<tr>
<td>20</td>
<td>0,180</td>
<td>0,181</td>
<td>0,191</td>
<td>0,191</td>
<td>0,178</td>
<td>0,179</td>
<td>0,179</td>
<td>0,180</td>
</tr>
<tr>
<td>50</td>
<td>0,138</td>
<td>0,139</td>
<td>0,138</td>
<td>0,138</td>
<td>0,137</td>
<td>0,138</td>
<td>0,131</td>
<td>0,133</td>
</tr>
<tr>
<td>100</td>
<td>0,117</td>
<td>0,117</td>
<td>0,117</td>
<td>0,117</td>
<td>0,117</td>
<td>0,117</td>
<td>0,113</td>
<td>0,113</td>
</tr>
</tbody>
</table>
Population III

<table>
<thead>
<tr>
<th>D</th>
<th>(1,2) post-strat.</th>
<th>(1,3) post-strat.</th>
<th>(2,1) post-strat.</th>
<th>(2,3) post-strat.</th>
<th>(3,1) post-strat.</th>
<th>(3,2) post-strat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.681</td>
<td>0.682</td>
<td>0.685</td>
<td>0.690</td>
<td>0.691</td>
<td>0.713</td>
</tr>
<tr>
<td>5</td>
<td>0.314</td>
<td>0.315</td>
<td>0.320</td>
<td>0.321</td>
<td>0.320</td>
<td>0.314</td>
</tr>
<tr>
<td>10</td>
<td>0.232</td>
<td>0.233</td>
<td>0.234</td>
<td>0.235</td>
<td>0.237</td>
<td>0.238</td>
</tr>
<tr>
<td>15</td>
<td>0.196</td>
<td>0.197</td>
<td>0.200</td>
<td>0.202</td>
<td>0.196</td>
<td>0.197</td>
</tr>
<tr>
<td>20</td>
<td>0.176</td>
<td>0.177</td>
<td>0.182</td>
<td>0.182</td>
<td>0.181</td>
<td>0.175</td>
</tr>
<tr>
<td>50</td>
<td>0.131</td>
<td>0.131</td>
<td>0.133</td>
<td>0.134</td>
<td>0.135</td>
<td>0.128</td>
</tr>
<tr>
<td>100</td>
<td>0.109</td>
<td>0.109</td>
<td>0.114</td>
<td>0.115</td>
<td>0.113</td>
<td>0.107</td>
</tr>
</tbody>
</table>

Population IV

<table>
<thead>
<tr>
<th>D</th>
<th>(1,2) post-strat.</th>
<th>(1,3) post-strat.</th>
<th>(2,1) post-strat.</th>
<th>(2,3) post-strat.</th>
<th>(3,1) post-strat.</th>
<th>(3,2) post-strat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.790</td>
<td>0.790</td>
<td>0.820</td>
<td>0.820</td>
<td>0.797</td>
<td>0.796</td>
</tr>
<tr>
<td>5</td>
<td>0.367</td>
<td>0.368</td>
<td>0.377</td>
<td>0.377</td>
<td>0.368</td>
<td>0.369</td>
</tr>
<tr>
<td>10</td>
<td>0.267</td>
<td>0.267</td>
<td>0.272</td>
<td>0.273</td>
<td>0.269</td>
<td>0.269</td>
</tr>
<tr>
<td>15</td>
<td>0.228</td>
<td>0.228</td>
<td>0.229</td>
<td>0.228</td>
<td>0.230</td>
<td>0.231</td>
</tr>
<tr>
<td>20</td>
<td>0.200</td>
<td>0.200</td>
<td>0.203</td>
<td>0.203</td>
<td>0.202</td>
<td>0.203</td>
</tr>
<tr>
<td>50</td>
<td>0.144</td>
<td>0.145</td>
<td>0.144</td>
<td>0.144</td>
<td>0.148</td>
<td>0.149</td>
</tr>
<tr>
<td>100</td>
<td>0.124</td>
<td>0.124</td>
<td>0.124</td>
<td>0.124</td>
<td>0.129</td>
<td>0.130</td>
</tr>
</tbody>
</table>

Population V

<table>
<thead>
<tr>
<th>D</th>
<th>(1,2) post-strat.</th>
<th>(1,3) post-strat.</th>
<th>(2,1) post-strat.</th>
<th>(2,3) post-strat.</th>
<th>(3,1) post-strat.</th>
<th>(3,2) post-strat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.734</td>
<td>0.734</td>
<td>0.700</td>
<td>0.702</td>
<td>0.748</td>
<td>0.748</td>
</tr>
<tr>
<td>5</td>
<td>0.344</td>
<td>0.345</td>
<td>0.331</td>
<td>0.332</td>
<td>0.352</td>
<td>0.353</td>
</tr>
<tr>
<td>10</td>
<td>0.258</td>
<td>0.258</td>
<td>0.241</td>
<td>0.241</td>
<td>0.264</td>
<td>0.265</td>
</tr>
<tr>
<td>15</td>
<td>0.219</td>
<td>0.219</td>
<td>0.205</td>
<td>0.205</td>
<td>0.225</td>
<td>0.226</td>
</tr>
<tr>
<td>20</td>
<td>0.195</td>
<td>0.196</td>
<td>0.186</td>
<td>0.186</td>
<td>0.202</td>
<td>0.202</td>
</tr>
<tr>
<td>50</td>
<td>0.141</td>
<td>0.141</td>
<td>0.138</td>
<td>0.139</td>
<td>0.146</td>
<td>0.147</td>
</tr>
</tbody>
</table>
Remark

The reader can find more details on algorithms used for generating populations and for calculations of estimators as well as more tables with results in Jakubowski and Bracha (2001).

4. Conclusions

The simulation study has given good results. The obtained results allow us to state that for many complex sampling surveys, similar to that realized by GUS, the Rao–Shao method (repeatedly grouped balanced half-samples method) is a very useful method of approximative variance estimation. The big advantage of this method is its simplicity. Therefore it may be recommended to use in practice, for example in GUS to such survey as household budget surveys, labour force survey and living condition survey.

Acknowledgement

The research was supported by the grant 1 H02B 008 17 from the Polish Committee for Scientific Research (KBN). The authors would like to thank Mr Janusz Raniszewski for computer calculation according to algorithms.

REFERENCES


P.C. MAHALANOBIS (1939), A sample survey of the acreage under jute in Bengal. Sankhya 4, pp. 511-531.


ON THE USE OF TRANSFORMED AUXILIARY VARIABLE IN THE ESTIMATION OF POPULATION MEAN IN TWO PHASE SAMPLING

G.N. Singh

ABSTRACT

The present investigation deals with the problem of estimating the population mean in two-phase sampling. New chain type estimators have been proposed which make use of known standard deviation, coefficient of skewness and coefficient of kurtosis of the second auxiliary character. The proposed estimators have been compared with two-phase ratio estimator and some other chain type estimators. Further generalizations have been made and its properties are discussed. The performances of the proposed estimators have been supported with a numerical illustration.

Key Words: Double sampling, chain ratio – type estimator, coefficient of variation, coefficient of skewness, coefficient of kurtosis, bias, mean squared error.

1. Introduction

The ratio method of estimation is the well known technique for estimating the population mean of a study character when the population mean of an auxiliary character is known and it is positively correlated with the study character. In the absence of the knowledge on the population mean of the auxiliary character we go for two-phase (double) sampling. The two-phase sampling happens to be a powerful and cost effective (economical) technique for obtaining the reliable estimate in first phase sample for the unknown parameters of the auxiliary character and hence has an eminent role to play in survey sampling, see Hidiroglou and Sarandal (1995, 98).

In order to construct an efficient estimator of the population mean of the auxiliary character in first-phase (preliminary) sample, Chand (1975) gave a technique of chaining another auxiliary character (which is highly correlated with

---

1 Department of Applied Mathematics, Indian School of Mines, Dhanbad-826 004, INDIA, E.mail: gnsingh_ism@yahoo.com
first auxiliary character but remotely correlated with the study character) with the first auxiliary character by using the ratio estimator in the first phase sample. The estimator is known as chain ratio-type estimator. Further, this work was extended by Kiregyera (1980, 1984), Singh and Singh (1991), Singh et al. (1994), Singh and Upadhyaya (1995), Singh et al. (2000), Upadhyaya and Singh (2001) and many others by proposing several chain-type ratio and regression estimators.

For making the estimates more precise, Searls (1964) used the coefficient of variation (CV) of study character at estimation stage. In practice this CV is seldom known. Motivated by Searls (1964) work, Sisodia and Dwivedi (1981) suggested a modified ratio estimator for population mean of study character by using the known CV of the auxiliary character. Later on Singh and Karan (1993), Upadhyaya and Singh (1999) proposed another ratio-type estimators with utilizing the known CV and coefficient of kurtosis (CK) of the auxiliary character. All these authors have used the CV and CK of auxiliary character in additive form to the sample and population means of the same character. It could be noticed that CV and CK are unit free constants, their additions may not be justified. Further, if CV and population mean of auxiliary character are known, standard deviation (SD) of auxiliary character is automatically known and the use of standard deviation in additive form is more justified, see Srivastava and Jhajj (1980).

Motivated by above points, in this work, an attempt has been made to utilize the information on known standard deviation (SD), coefficient of skewness (CS) and coefficient of kurtosis (CK) of the second auxiliary character through a most suitable transformation for estimating the population mean of auxiliary character more precisely in the first-phase (preliminary) sample. Three new chain-type estimators for the population mean of the study character have been proposed. The proposed estimators have been compared with two-phase ratio estimator and some other chain-type estimators. Further, generalizations have been made and its properties are discussed. The performances of the proposed estimators have been supported with a numerical illustration.

2. Proposed Estimators

Let \( U = (u_1, u_2, \ldots, u_N) \) be the finite population of \( N \) units, \( y \) and \( x \) be the characters under study and first auxiliary character respectively. It is assumed that \( y \) and \( x \) are highly positively correlated. Let \( y_k > 0 \) and \( x_k > 0 \) be the values of \( y \) and \( x \) for the \( k \)-th ( \( k = 1, 2, \ldots, N \) ) unit in the population. From the population \( U \), a simple random sample of size \( n \) is drawn without replacement. Let \( (\bar{Y}, \bar{X}) \) and \( (\bar{y}, \bar{x}) \) be the population means and sample means of the respective variates \( y \) and \( x \) respectively. The classical ratio estimator for \( \bar{Y} \) is defined as

\[
\bar{y}_r = \frac{\bar{Y}}{\bar{X}} \bar{x}
\]  

(1)
If $\bar{X}$ is not known, we estimate $\bar{Y}$ by two-phase ratio estimator

$$\bar{Y}_{rd} = \frac{\bar{Y}}{\bar{X}} \bar{x}'$$

(2)

where $\bar{x}'$ is the sample mean of $x$ based on the first-phase (preliminary) sample of size $n' > n$.

The way in which the estimate of $\bar{Y}$ is improved using the auxiliary information on $x$ can also be extended to improve the estimator of $\bar{X}$ in the first-phase sample, if another auxiliary variable $z$ closely related to $x$ but remotely related to $y$ is used. Thus assuming that the population mean $\bar{Z}$ of the variable $z$ is known, Chand (1975) proposed a chain-type ratio estimator as

$$\bar{y}_c = \frac{\bar{Y}}{\bar{X}} \bar{x}_{rd},$$

(3)

with $\bar{x}_{rd} = \frac{\bar{x}'}{\bar{z}'}$, where $\bar{z}'$ is the mean of the variable $z$ in the first-phase sample of size $n'$.

Utilizing the known coefficient of variation of $z$ Singh and Upadhyaya (1995) considered a modified chain-type ratio estimator as

$$\bar{y}_{mc} = \frac{\bar{Y}}{\bar{X}} \bar{x}' \left[ \frac{\bar{Z} + C_z}{\bar{z} + C_z} \right]$$

(4)

where $C_z$ is the known coefficient of variation of the variable $z$.

In many situations the values of the auxiliary variate may be available for each unit in the population, for instance, see Das and Tripathi (1981). In such situations knowledge on $\bar{Z}$, $C_z$, $\beta_1(z)$ (coefficient of skewness), $\beta_2(z)$ (coefficient of kurtosis) and possibly on some other parameters may be utilized. Regarding the availability of information on $C_z$, $\beta_1(z)$ and $\beta_2(z)$, the researchers may be referred to Searls (1964), Sen (1978), Murthy (1967, pp. 96-99), Singh et al. (1973) and Searls and Intarpanich (1990).

Using the known coefficient of variation $C_z$ and known coefficient of kurtosis $\beta_2(z)$ of the second auxiliary character $z$ Upadhyaya and Singh (2001) considered the following estimators for $\bar{Y}$
If the population mean and coefficient of variation of the second auxiliary character is known, the standard deviation $\sigma_z$ is automatically known and it is more meaningful to use the $\sigma_z$ in addition to $C_z$, see Srivastava and Jhajj (1980). Further, $C_z$, $\beta_1(z)$ and $\beta_2(z)$ are the unit free constants, their use in additive form is not much justified. Motivated with the above justifications and utilizing the known values of $\sigma_z$, $\beta_1(z)$ and $\beta_2(z)$, we suggest the following three more reasonable transformations for $z$ and corresponding estimators for $\bar{Y}$ which are as follows:

Let $v_{ik} = \alpha_i z_k + \sigma_z$, (i= 1, 2, 3 and k=1, 2, . . ., N) so that $\bar{v}_i = \alpha_i \bar{z} + \sigma_z$ is the sample mean of the transformed variate $v$ in the first-phase sample and $\bar{V}_i = \alpha_i \bar{Z} + \sigma_z$ is the corresponding population mean, where $\alpha_1=1$, $\alpha_2 = \beta_1(z)$ and $\alpha_3 = \beta_2(z)$ are the known values of $\alpha_i$’s for i=1, 2, 3. Subsequently, the following three new chain-type estimators are considered as

$$T_{s1} = \bar{Y} - \bar{\bar{v}} + \bar{\bar{V}}_1, \text{ for } i = 1, \alpha_1=1$$

$$T_{s2} = \bar{Y} - \bar{\bar{v}} + \bar{\bar{V}}_2, \text{ for } i = 2, \alpha_2 = \beta_1(z)$$

and

$$T_{s3} = \bar{Y} - \bar{\bar{v}} + \bar{\bar{V}}_3, \text{ for } i = 3, \alpha_3 = \beta_2(z)$$
Estimators, proposed in (7)-(9) can be re-written as a sequence of estimators in the following form:

\[
T_{si} = \frac{\bar{y} - \bar{x}}{x} \left[ \frac{\alpha_i \bar{Z} + \sigma_z}{\alpha_i \bar{Z} + \sigma_z} \right] = \frac{\bar{y} - \bar{x}}{x} \bar{v}_i, \quad (i = 1, 2, 3)
\]

(10)

3. Bias and Mean Squared Errors (M.S.E.’s) of the Proposed Sequence of Estimators \( T_{si} \) (\( i = 1, 2, 3 \))

As the proposed sequence of estimators \( T_{si} \) (\( i = 1, 2, 3 \)) in (10) are biased for \( \bar{Y} \), their biases and mean squared errors (m.s.e.’s) have been obtained up to the first order of approximations under the following transformations:

\[
\bar{y} = \bar{Y}(1 + e_1), \quad \bar{x} = \bar{X}(1 + e_2), \quad \bar{z} = \bar{Z}(1 + e_4) \quad \text{with} \quad E(e_i) = 0, \quad \text{for} \quad j = 1, 2, 3, 4.
\]

The proposed sequence of estimators \( T_{si}(i = 1, 2, 3) \) then become

\[
T_{si} = (1 + e_1)(1 + e_2)^{-1}(1 + e_3)(1 + \phi_4)^{-1}, \quad (i = 1, 2, 3)
\]

(11)

where \( \phi_1 = \frac{\bar{Z}}{(Z + \sigma_z)}, \quad \phi_2 = \frac{\beta_1(z)\bar{Z}}{(\beta_1(z)\bar{Z} + \sigma_z)} \quad \text{and} \quad \phi_3 = \frac{\beta_2(z)\bar{Z}}{(\beta_2(z)\bar{Z} + \sigma_z)} \).

Realizing that for \( i = 1, 2, 3 \) \( \phi_i < 1 \) and assuming \( |\phi_i e_4| < 1 \) and \( |e_2| < 1 \), we expand the terms of (11) and collecting the terms up to the first order of approximation, we have the following results:

**Theorem 1:** The bias \( B(.) \) and mean squared error (m.s.e.) \( M(.) \) of the proposed sequence of estimators \( Tsi \) (\( i = 1, 2, 3 \)), to the terms of order \( o(n^{-1}) \) are given by

\[
B(T_{si}) = E(T_{si} - \bar{Y}) = \bar{Y} \left[ f_1(c_x^2 - \rho_{xy} c_y c_x) + f_2(\phi_1^2 C_z^2 - \phi_1 \rho_{yz} C_y C_z) \right]
\]

and

\[
M(T_{si}) = E(T_{si} - \bar{Y})^2 = \bar{Y}^2 \left[ f_1 C_y^2 + f_2(\phi_1^2 C_z^2 - \phi_1 \rho_{yz} C_y C_z) + f_3(C_y^2 - \rho_{yx} C_y C_x) \right]
\]

(12)

where \( \rho_{yx}, \rho_{yz} \) and \( \rho_{zx} \) are the population coefficients of correlation between the variables used in suffices, \( C_t \) is the coefficient of variation of the variable \( t \) (\( t = x, y, z \)) and \( f_1 = \left( \frac{1}{n} - \frac{1}{N} \right), \quad f_2 = \left( \frac{1}{n'} - \frac{1}{N} \right), \quad f_3 = \left( \frac{1}{n'} - \frac{1}{n} \right). \)
Corollary 1: It is obvious that biases and m.s.e.’s of $T_{s1}$, $T_{s2}$ and $T_{s3}$ defined in (7) – (9) can be obtained by substituting the values of $\phi_i$ (i = 1,2,3) in (12) and (13) respectively.

4. Efficiency Comparisons of $T_{si}$ (i =1, 2, 3)

In this section, the conditions for which the proposed sequence of estimators are better than $\bar{y}_{rd}$, $\bar{y}_{c}$, $\bar{y}_{mc}$, $\bar{y}_{us1}$, $\bar{y}_{us2}$ have been obtained. The m.s.e.’s of these estimators up to the order $o(n^{-1})$ are derived as

$$M(\bar{y}_{rd}) = \bar{Y}^2 \left[ f_1 C_y^2 + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) \right]$$ \hspace{1cm} (14)

$$M(\bar{y}_{c}) = \bar{Y}^2 \left[ f_1 C_y^2 + f_2 (C_z^2 - 2 \rho_{yz} C_y C_z) + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) \right]$$ \hspace{1cm} (15)

$$M(\bar{y}_{mc}) = \bar{Y}^2 \left[ f_1 C_y^2 + f_2 (\theta^2 C_z^2 - 2 \rho_{yz} C_y C_z) + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) \right]$$ \hspace{1cm} (16)

$$M(\bar{y}_{us1}) = \bar{Y}^2 \left[ f_1 C_y^2 + f_2 (P^2 C_z^2 - 2 \rho_{yz} C_y C_z) + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) \right]$$ \hspace{1cm} (17)

and

$$M(\bar{y}_{us2}) = \bar{Y}^2 \left[ f_1 C_y^2 + f_2 (Q^2 C_z^2 - 2 \rho_{yz} C_y C_z) + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) \right]$$ \hspace{1cm} (18)

where $\theta = \frac{Z}{Z + C_z}$, $P = \frac{\beta_z(z) \bar{Z}}{\beta_z(z) \bar{Z} + C_z}$ and $Q = \frac{C_z \bar{Z}}{C_z \bar{Z} + \beta_z(z)}$

4.1. Comparison of $T_{si}$ (i =1,2,3) with $\bar{y}_{rd}$ and $\bar{y}_{c}$

It is obvious from (13) and (14) that $M(T_{si}) < M(\bar{y}_{rd})$ if

$$\rho_{yz} > \frac{1}{2} \frac{C_z}{C_y} \phi_i$$ \hspace{1cm} (19)

It follows from (13) and (15) that the sequence of estimators $T_{si}$ is more efficient than $\bar{y}_{c}$ if

$$\rho_{yz} < \frac{1}{2} \frac{C_z}{C_y} (1 + \phi_i)$$ \hspace{1cm} (20)

subsequently the results in (19) and (20) are combined in the following theorem:
Theorem 2: The range of $\rho_{yz}$ for which $T_{si}$ ($i = 1, 2, 3$) is better than $\bar{y}_{nl}$ and $\bar{y}_c$ is

$$\frac{1}{2} \frac{C_z}{C_y} \phi_i < \rho_{yz} < \frac{1}{2} \frac{C_z}{C_y} (1 + \phi_i)$$

(21)

Corollary 2: Suitable ranges of $\rho_{yz}$ for the estimators $T_{s1}$, $T_{s2}$ and $T_{s3}$ can be derived by substituting the respective values of $\phi_i$ ($i = 1, 2, 3$) in (21).

4.2 Comparison of $T_{si}$($i=1, 2, 3$) with $\bar{y}_{mc}$, $\bar{y}_{us1}$ and $\bar{y}_{us2}$

From (13), (16), (17) and (18) it could be concluded that

(i) $T_{si}$ is better than $\bar{y}_{mc}$ if $\rho_{yz} < \frac{1}{2} \frac{C_z}{C_y} (\theta + \phi_i)$, ($i = 1,2,3$)

(22)

(ii) $T_{si}$ is more precise than $\bar{y}_{us1}$ if $\rho_{yz} < \frac{1}{2} \frac{C_z}{C_y} (P + \phi_i)$, ($i = 1,2,3$)

(23)

and

(iii) $T_{si}$ dominates $\bar{y}_{us2}$ if $\rho_{yz} < \frac{1}{2} \frac{C_z}{C_y} (\theta + \phi_i)$, ($i = 1,2,3$)

(24)

Remark: When the auxiliary character $x$ and the second auxiliary character $z$ is negatively correlated, we may consider the following sequence of chain-type product estimators as

$$T_{si}^* = \bar{y} \frac{\alpha_i Z + \sigma_z}{x \alpha_i Z + \sigma_z} = \frac{\alpha_i Z^* + \sigma_z^*}{x \alpha_i Z + \sigma_z}, \quad (i = 1,2,3)$$

(25)

5. Sequence of Generalized Class of Estimators of $\bar{Y}$

Motivated by Srivastava (1967), we present a sequence of generalized class of estimators of $\bar{Y}$ as

$$T_{gpi}^* = \bar{y} \left[ \frac{\alpha_i Z + \sigma_z}{x \alpha_i Z + \sigma_z} \right]^g \bar{y} \left[ \frac{\alpha_i Z^* + \sigma_z^*}{x \alpha_i Z + \sigma_z} \right]^g, \quad (i = 1,2,3)$$

(26)
where \( g \) is a suitably chosen scalar. For \( g = 1 \), \( T_{gsi} \) reduces to \( T_{si} \) while for \( g = -1 \), \( T_{gsi} \) yields the estimator \( T_{si}^* \) and when \( g = 0 \), \( T_{gsi} \) boils down to the conventional two-phase ratio estimator \( \bar{y}_{rd} \).

The bias and m.s.e. of the estimator \( T_{gsi} \) to the first degree of approximations are respectively given by

\[
B(T_{gsi}) = \bar{y} \left[ f_3 (C_x^2 - \rho_{yx} C_y C_x) + f_2 \left[ g(1 + 1) \phi_i^2 C_z^2 / 2 - g \phi_i \rho_{yz} C_y C_z \right] \right] \quad (27)
\]

and

\[
M(T_{gsi}) = \bar{y} \left[ f_1 C_y^2 + f_2 (g^2 \phi_i^2 C_z^2 - 2g \phi_i \rho_{yx} C_y C_z) + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) \right] \quad (28)
\]

The biases and m.s.e.’s of \( T_{si} \), \( T_{si}^* \) and \( \bar{y}_{rd} \) can easily be obtained from (27) and (28) respectively by putting suitable values of \( g \).

5.1 Optimum value of \( g \) and Minimum M.S.E. of \( T_{gsi} \) (i = 1,2,3)

**Theorem 3:** The \( M(T_{gsi}) \) in (28) is minimized for

\[
g_{opt} = \frac{\rho_{yx} C_y}{\phi_i C_z} \quad (29)
\]

and the minimum m.s.e. of the sequence of generalized class of estimators \( T_{gsi} \) (i = 1, 2, 3) is given by

\[
\min M(T_{gsi}) = \bar{y} \left[ f_1 C_y^2 + f_3 (C_x^2 - 2 \rho_{yx} C_y C_x) - f_2 \rho_{yz} C_z^2 \right] \quad \forall i=1, 2, 3 \quad (30)
\]

**Remark:** It can be easily shown that \( T_{gsi} \) (i = 1,2,3) is always better than \( \bar{y}_{rd} \), \( \bar{y}_{ce} \), \( \bar{y}_{mc} \), \( \bar{y}_{us1} \), and \( \bar{y}_{us2} \) at the optimum value of \( g \) given in (29). The optimum value of \( g \) i.e. \( g_{opt} = \frac{\rho_{yx} C_y}{\phi_i C_z} \) can be guessed quite accurately from the past data or experience gathered in due course of time.

6. Efficiency Comparisons of \( T_{gsi} \) (i = 1,2,3)

It is important to investigate the situations under which the proposed sequence of generalized class of estimators \( T_{gsi} \) (i = 1,2,3) is preferable over the estimators discussed in the present work. We present below the ranges of \( g \) for
which $T_{gsi}$ ($i = 1, 2, 3$) is preferable over the estimators $\bar{y}_{rd}$, $\bar{y}_c$, $\bar{y}_{mc}$, $\bar{y}_{us1}$ and $\bar{y}_{us2}$.

(i) It is obvious from (28) and (14) that

$$M(T_{gsi}) \leq M(\bar{y}_{rd}) \text{ if } g \leq 2\rho_{yz} \frac{C_z}{\phi_i C_z}$$

(ii) A comparison of (28) and (15) reveals that $T_{gsi}$ is preferable over $\bar{y}_c$ if

$$\frac{1}{\phi_i} \leq g \leq 2\rho_{yz} \frac{C_z}{\phi_i C_z} - \frac{1}{\phi_i}$$

(iii) From (28) and (16) we find that $M(T_{gsi}) \leq M(\bar{y}_{mc})$ if

$$\frac{\theta}{\phi_i} \leq g \leq 2\rho_{yz} \frac{C_z}{\phi_i C_z} - \frac{\theta}{\phi_i}$$

(iv) From (28) and (17) we observe that $M(T_{gsi}) \leq M(\bar{y}_{us1})$ if

$$\frac{P}{\phi_i} \leq g \leq 2\rho_{yz} \frac{C_z}{\phi_i C_z} - \frac{P}{\phi_i}$$

and

(v) Finally, from (28) and (18) we conclude that $T_{gsi}$ is better than $\bar{y}_{us2}$ if

$$\frac{Q}{\phi_i} \leq g \leq 2\rho_{yz} \frac{C_z}{\phi_i C_z} - \frac{Q}{\phi_i}$$

7. Numerical Illustrations

We consider the data used by Anderson (1958) to demonstrate what we have discussed earlier. 25 families have been observed for the following three variables.

$y$: Head length of second son
$x$: Head length of first son
$z$: Head breadth of first son
\( \bar{Y} = 183.84, \bar{X} = 185.72, \bar{Z} = 151.12, \sigma_z = 7.2240, C_y = 0.0546, C_x = 0.0526, C_z = 0.0488, \rho_{xy} = 0.7108, \rho_{yz} = 0.6932, \rho_{xz} = 0.7346, \beta_1(z) = 0.0002, \beta_2(z) = 2.6519. \) Consider \( n = 7 \) and \( n' = 10 \)

Let \( E = \frac{M(T)}{M(T_{gsi})} \times 100 \) be the percent efficiency of the proposed sequence of generalized estimator \( T_{gsi} (i = 1,2,3) \) under optimum condition with respect to the estimator \( T \). The m.s.e.'s of the different estimators and respective values of \( E \) are presented in Table 1.

### Table 1

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( T_{esi} ) for ( i = 1,2,3 )</th>
<th>( T_{s1} )</th>
<th>( T_{s2} )</th>
<th>( T_{s3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>5.5538</td>
<td>5.7075</td>
<td>8.4287</td>
<td>5.7594</td>
</tr>
<tr>
<td>E</td>
<td>102.7678</td>
<td>151.7649</td>
<td>103.7014</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( \bar{y}_{rd} )</th>
<th>( \bar{y}_c )</th>
<th>( \bar{y}_{mc} )</th>
<th>( \bar{y}_{w1} )</th>
<th>( \bar{y}_{w2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>8.4599</td>
<td>5.7962</td>
<td>5.7955</td>
<td>5.7959</td>
<td>5.5617</td>
</tr>
<tr>
<td>E</td>
<td>152.3255</td>
<td>104.3639</td>
<td>104.3513</td>
<td>104.3591</td>
<td>100.1423</td>
</tr>
</tbody>
</table>

From Table 1 we observe that the proposed sequence of generalized class of estimators \( T_{gsi} \) is preferable over all the considered estimators under optimum condition. The estimators \( T_{s1} \) and \( T_{s3} \) are either better or approximately equally efficient than the other estimators such as \( \bar{y}_{rd}, \bar{y}_c, \bar{y}_{mc}, \bar{y}_{w1}, \) and \( \bar{y}_{w2} \).

Therefore, the transformations have been considered in this work are more justifiable in compare with the previous works of the similar nature. The estimator \( T_{s2} \) is not performing well, because the \( \beta_1(z) \) of this population is approximately zero (symmetrical population) while it could be observed that \( T_{s2} \) will perform well for the skewed populations.

### Conclusions

The transformations which are considered in this work are more reasonable and from the above results, we can conclude that the estimators which are proposed in this paper are either better than other estimators or approximately
equally efficient. Therefore, the use of these estimators may be recommended by the survey statisticians.

REFERENCES


ON SOME EQUIVALENT DEFINITIONS OF SIMPLE RANDOM SAMPLING AND THEORETICAL JUSTIFICATION OF VARIOUS SELECTION METHODS THEREBY

G.C. Tikkiwal

ABSTRACT

Tikkiwal, B.D. (1984) gave three equivalent definitions for each of the two sampling schemes: Simple Random Sampling with and without replacement. This paper first gives a rigorous proof of the equivalence of these definitions in each case. Making use of these three equivalent definitions, it then gives theoretical justification of the selection of simple random samples through various methods in vogue; as there has been no rigorous justifications for the use of such methods including even the earliest urn method.

Keywords: Sampling design, sampling scheme, random number, simple random sample, simple random sampling without replacement (SRSWOR), simple random sampling with replacement (SRSWR), compound experiment, draw as a success.

1. Introduction

A simple random sample by definition is a sample selected from a population through either of the two sampling schemes: simple random sampling without replacement (SRSWOR) and simple random sampling with replacement (SRSWR). Before the use of Tippett's random number in 1927, the urn method or its variant the chit method was used for drawing such samples. Since the units of a finite population together with their variate values can be represented by the duly marked balls of an urn, the urn method consisted in drawing the balls one by one, with or without replacement as the case may be, till we get a sample of fixed size n. Currently the practice is to use random numbers for drawing such samples.

A random number table was first given by Tippett (1927). Since then we have random number tables available from other sources such as one of Kendall

1 Dept. of Mathematics and Statistics, J.N.V. University, Jodhpur-342001, India.
and Smith (1938) and the other of Rand Corporation (1955). The various methods
of selecting simple random samples using the random numbers tables are in
practice [Murthy (1967), Cochran (1977)]. These methods of selection consist of
associating one or several random numbers to population units for the purpose of
selection.

In the method of associating one random number, many random numbers get
rejected greater than the population size. This difficulty can be overcome to a
large extent by associating with each unit of the population an equal number of
random numbers and then using remainder or quotient approach for the selection
of the sample.

It is believed that the methods, such as ones described above lead to simple
random samples. However, there appears to be no theoretical justification of this
belief in the literature, including even of the urn method. This we provide in
Section 4, making use of equivalent definitions of simple random sampling with
Before we discuss the equivalent definitions of simple random sampling, we
present certain preliminaries for the same.

2. Preliminaries

Let us consider a finite population consisting of \(N\) \((< \infty)\) distinguishable
units \((u_1, \ldots, u_i, \ldots, u_N)\), labeled \(i = 1,2,\ldots,N\). For simplicity, we let the \(i\)-th unit be
represented by its label \(i\). Thus, we denote finite population as \(\pi : (1,2,\ldots,N)\). A
sample \(s_t = (i_1, i_2,\ldots,i_n)\) from \(\pi\) is a finite ordered sequence of units from \(\pi\); where
\(i_r\) denotes the event of occurrence of some particular unit drawn at the \(r\)-th draw
for \(r = 1,2,\ldots,n\). Here the units need not be distinct from each other; since they
represent the units drawn with or without replacement in \(n\) consecutive draws.
Further, let \(S = \{ s_t \}\) denote the set of all possible samples \(s_t\). An ordered
sampling design [ cf. Cassel et . al (1977)] is a pair \((S,P)\) where \(P\) is a probability
measure defined over \(S\) such that \(P(s_t) \geq 0\) for all \(s_t \in S\) and \(\sum_{s_t \in S} P(s_t) = 1\).

Instead of considering an ordered sequence \(s_t\), if we take \(s_t\) as an unordered
sequence; then the design is called an unordered sampling design. Any
mechanism of selecting a sample \(s_t\) resulting in probability \(P(s_t)\) for \(s_t \in S\) as
above is called a sampling scheme. Such a sampling scheme always exist in view
of the result due to Rao (1962), that corresponding to any sampling design there
exists a sampling scheme in which units are drawn one by one.

The classical definition of the two sampling schemes, SRSWR and
SRSWOR, consists of drawing units one by one, with or without replacement as
the case may be, till we get a sample of size \(n\) in such a manner that at each draw
the available units get equal chance of selection. That, this happens in the urn
method, for drawing simple random samples, is intuitively clear. However, this is
not a rigorous way to look at the things. But when the urn method is looked upon
as an urn model, one immediately appreciates that drawing of n units from N units
of a given population, through urn device, can be regarded as n random
experiments, mutually independent in case of SRSWR and mutually dependent in
case of SRSWOR [cf. Kolmogorov, A.N. (1956)]. The sample of size n is then the
result of the random experiment E, a compound one of these n random
experiments. When the classical definition is interpreted in terms of this random
experiment E, then, in case of without replacement, it means that
\[
P \{(i, r) | i_1, i_2, \ldots, i_{r-1} \neq i\} = \frac{1}{N - r + 1} \tag{2.1}
\]
for all i and r and for all the sets of \((i_1, i_2, \ldots, i_{r-1} \neq i)\); where
\((i, r)\) denotes the event that the i-th unit occurs at the r-th draw,
for \(i=1,2,\ldots, N\) and \(r=1,2,\ldots, n\). \tag{2.2}
and \((i_1, i_2, \ldots, i_{r-1} \neq i)\) denote a set of \((r-1)\) events in which some \((r-1)\) different
units, other than the i-th unit, occur at the first \((r-1)\) draws. Here the symbols ‘\#’
is used for ‘not equal to’.
Using the fundamental principle of counting, there are
\[
S(r) = \left\{ \binom{N - 1}{r - 1} (r - 1)! \right\} \left\{ \binom{N - r}{n - r} (n - r)! \right\} = \binom{N - 1}{n - 1} (n - 1)! \tag{2.3}
\]
such samples in number, out of the total number of possible samples \(\binom{N}{n} n!\), in
which i-th unit occurs at the r-th draw.
Then,
\[
\{(\tilde{i}, r'), r' \leq r - 1\} = U \{(i_1, i_2, \ldots, i_{r-1} \neq i)\}
\tag{2.4}
\]
the union being over all \(S(r)\) sets; where, \(\{(\tilde{i}, r'), r' \leq r - 1\}\) represents an event in
which i-th unit does not occur at any of the previous \((r-1)\) draws. Also, let
\[
\{(i_1, 1) | i_0 = 1, i_1, i_2, \ldots, i_{r-1} \neq i\} \equiv \{(i, 1) | \tilde{(i, 0)}\} \equiv (i_1, 1) \tag{2.5}
\]
Now we state and prove the following lemma, which provides us a definition
of SRSWOR alternate to the classical definition in (2.1).

**Lemma 2.1.** In case of without replacement,
\[
P \{(i, r) | i_1, i_2, \ldots, i_{r-1} \neq i\} = \frac{1}{N - r + 1}
\]
for all i and r and for all the S(r) sets \( (i_1, i_2, \ldots, i_{r-1} \neq i) \); if and only if

\[
P\{(i, r) | i \} , r' \leq r - 1 \} = \frac{1}{N - r + 1}
\]

for all i and r; and \( P\{(i, r) | i, i_1, i_2, \ldots, i_{r-1} \neq i \} \) is same for all the S(r) sets \( (i_1, i_2, \ldots, i_{r-1} \neq i) \).

**Proof.** Let

\[
P\{(i, r) | i_1, i_2, \ldots, i_{r-1} \neq i \} = \frac{1}{N - r + 1}
\]

Then,

\[
P\{(i, r) | i \} , r' \leq r - 1 \} = \frac{P\{(i, r), (i, r - 1), \ldots, (i, l) \}}{P\{(i, r - 1), \ldots, (i, l) \}}
\]

\[
= \sum_{i_1, i_2, \ldots, i_{r-1} \neq i} \frac{P\{(i, r), i_{r-1}, \ldots, i_1 \}}{P(i_1, \ldots, i_{r-1})}
\]

Now,

\[
P\{(i, r), i_{r-1}, \ldots, i_1 \} = P\{(i, r) | i_{r-1}, \ldots, i_1 \} P(i_{r-1} | i_{r-2}, \ldots, i_1) \ldots P(i_2 | i_1) P(i_1)
\]

\[
= \left( \frac{1}{N - r + 1} \right) \left( \frac{1}{N - r + 2} \right) \ldots \left( \frac{1}{N} \right)
\]

Also,

\[
P(i_1, \ldots, i_{r-1}) = P(i_{r-1} | i_{r-2}, \ldots, i_1) P(i_{r-2} | i_{r-3}, \ldots, i_1) \ldots P(i_2 | i_1) P(i_1)
\]

\[
= \left( \frac{1}{N - r + 2} \right) \left( \frac{1}{N - r + 3} \right) \ldots \left( \frac{1}{N} \right)
\]

Therefore,

\[
P\{(i, r) | i \} , r' \leq r - 1 \} = \frac{1}{N - r + 1}
\]

Now, let

\[
P\{(i, r) | i \} , r' \leq r - 1 \} = \frac{1}{N - r + 1}
\]

Since, from above
\begin{align*}
P\{ (i, r) \mid (\tilde{i}, r'), r' \leq r - 1 \} & = \sum_{i_1, \ldots, i_{r-1} \neq i} P\{(i, r) \mid i_1, \ldots, i_{r-1} \} P(i_1, \ldots, i_{r-1}) \\
& = \frac{\sum_{i_1, \ldots, i_{r-1} \neq i} P(i_1, \ldots, i_{r-1})}{\sum_{i_1, \ldots, i_{r-1} \neq i} P(i_1, \ldots, i_{r-1})}
\end{align*}

Since, \( P\{ (i, r) \mid i_1, \ldots, i_{r-1} \neq i \} \) is same for all the sets of \((i_1, \ldots, i_{r-1} \neq i)\);

\[
P\{ (i, r) \mid i_1, \ldots, i_{r-1} \neq i \} = P\{ (i, r) \mid (\tilde{i}, r'), r' \leq r - 1 \} = \frac{1}{N-r+1}
\]

Thus, the converse is also true.

In case of with replacement, the classical definition, when interpreted in terms of the said random experiment \( E \), gives

\[
P \{ (i, r) \mid i_1, i_2, \ldots, i_{r-1} \} = \frac{1}{N}
\]

(2.6)

for all \( i \) and \( r \) and for all the sets of \((i_1, i_2, \ldots, i_{r-1})\) with the proviso that \( \{i, 1\}|i_0) \equiv (i, 1) \), where \((i_1, i_2, \ldots, i_{r-1})\) denote a set of \((r-1)\) events in which some \((r-1)\) units, not necessarily different from each other and not necessarily different from \( i \) th unit, occur at the first \((r-1)\) draws.

In the following section, we discuss other equivalent definitions of simple random sampling, with and without replacement, including the one provided by Lemma 2.1, which is an alternate to the classical definition in (2.1) as mentioned earlier.

### 3. The Equivalent Definitions of Simple Random Sampling (SRS)

Let there be a finite population \( \pi : (1, 2, \ldots, N) \) of \( N \) distinguishable units with \( 'x' \) denoting some variate under study and \( x_i \) the variate value of the \( i \)-th unit \( u_i \) of the population for \( i=1,2,\ldots,N \). In order to estimate the population total \( T \) or some other parameters; we draw a sample of size \( n \) with unit by unit draw mechanisms, with or without replacement as the case may be. With these observations, we now state below the following alternate definitions of SRSWOR and SRSWR due to Tikkiwal, B.D. (1984).

**Definition 1.**

(a) For SRSWOR

In a unit by unit draw without replacement, if
\begin{equation}
P\{\{i, r\}|\{\tilde{i}, r'\}, r' \leq r - 1\} = \frac{1}{N - r + 1}
\end{equation}

for all \(i\) and \(r\), and \(P\{(i, r)|i_1, \ldots, i_{r-1} \neq i\} = \frac{1}{N - r + 1}\) is same for all the \(S(r)\) sets \((i_1, i_2, \ldots, i_{r-1} \neq i)\); then it is said to be simple random sampling without replacement (SRSWOR).

The latter condition in the above definition does not necessarily mean that the conditional probability of \((i, r)\) given \((i_1, i_2, \ldots, i_{r-1} \neq i)\) is independent of \(i\) for all \(S(r)\) sets, as happens in the classical definition provided in (2.1) of Section 2.

(b) For SRSWR
In a unit by unit draw with replacement, if

\begin{equation}
P\{(i, r)|i_1, i_2, \ldots, i_{r-1}\} = \frac{1}{N}
\end{equation}

for all \(i\) and \(r\) and for all the \(S(r)\) sets \((i_1, i_2, \ldots, i_{r-1})\) with the proviso that \(\{(i, 1)|i\neq 0\} \equiv (i, 1)\); then it is said to be simple random sampling with replacement (SRSWR).

Definition 2.

(a) For SRSWOR
In a unit by unit draw without replacement, if

\begin{equation}
P(i, r) = \frac{1}{N}
\end{equation}

for all \(i\) and \(r\), and \(P\{(i, r)|i_1, \ldots, i_{r-1} \neq i\} = \frac{1}{N - r + 1}\) is same for all the \(S(r)\) sets \((i_1, i_2, \ldots, i_{r-1} \neq i)\); then it is said to be SRSWOR. It may be noted that the latter condition here is the same as the latter condition occurring in Definition 1.

(b) For SRSWR
In a unit by unit draw with replacement, if

\begin{equation}
P(i, r) = \frac{1}{N}
\end{equation}

for all \(i\) and \(r\); then it is said to be SRSWR.

Definition 3.

(a) For SRSWOR
In a unit by unit draw without replacement, if probability of an ordered sample \(s_t\) is given by

\begin{equation}
P(s_t) = \frac{1}{\binom{N}{n} n!}
\end{equation}
for \( t = 1, 2, \ldots, \left(\begin{array}{c} N \\n \end{array}\right) n! \) giving all the ordered samples \( s_t \), in the compound random experiment \( E \), where \( s_t \) consists of \( n \) distinct units of the events \( i_r \) for \( r = 1, 2, \ldots, n \); then it is said to be SRSWOR.

(b) For SRSWR
In a unit by unit draw with replacement, if probability of the ordered sample \( s_t \) is given by

\[
P(s_t) = \frac{1}{N^n},
\]

for \( t = 1, \ldots, N^n \) giving all the ordered samples \( s_t \), in the compound random experiment \( E \), where \( s_t \) consists of \( n \) units, not necessarily distinct of the events \( i_r \) for \( r = 1, \ldots, n \); than it is said to be SRSWR.

In order to prove the equivalence of the three definitions 1(a), 2(a) and 3(a) of SRSWOR, we first present the following lemmas alongwith their proof.

**Lemma 3.1.** For SRSWOR

\[
P(i, r) = \frac{1}{N} \iff P\{ (i, r) \mid (\tilde{i}, r') \}, r' \leq r - 1\} = \frac{1}{N - r + 1}
\]

for all \( i \) and \( r \).

**Proof.** We first prove

\[
P(i, r) = \frac{1}{N} \Rightarrow P\{ (i, r) \mid (\tilde{i}, r') \}, r' \leq r - 1\} = \frac{1}{N - r + 1}
\]

for all \( i \) and \( r \).

The above result is obviously true for \( r = 1 \). For \( r = 2 \),

\[
P\{ (i, 2) \mid (\tilde{i}, 1) \} = \frac{P(i, 2)}{P(\tilde{i}, 1)} = \frac{P(i, 2)}{1 - P(i, 1)} = \frac{1}{N - 1}
\]

So (3.2) is true for \( r = 2 \). Let this now be true for \( r = 1, \ldots, s \). Then

\[
P\{ (i, s + 1) \mid (\tilde{i}, r') \}, r' \leq s\} = \frac{P(i, s + 1)}{P(\tilde{i}, r'), r' \leq s}
\]

But

\[
P\{ (i, r') \}, r' \leq s\} = P\{ (\tilde{i}, r') \}, r' \leq s - 1\} P\{ (\tilde{i}, s) \mid (\tilde{i}, r') \}, r' \leq s - 1\}
\]

\[
= P(\tilde{i}, 1)P\{ (\tilde{i}, 2) \mid (\tilde{i}, 1) \} \ldots P\{ (\tilde{i}, s) \mid (\tilde{i}, r') \}, r' \leq s - 1\} = \frac{N - s}{N}
\]
Therefore,

\[
P\{(i, s+1)| (\hat{i}, r'), r' \leq s\} = \frac{1}{N-s}
\]

showing that the result is true for \(r=s+1\). Thus, by induction, it is true in general.

We now prove the converse part of the lemma, i.e.

\[
P\{(i, r)| (\hat{i}, r'), r' \leq r-1\} = \frac{1}{N-r+1} \Rightarrow P(i, r) = \frac{1}{N}
\]

for all \(i\) and \(r\).

We have

\[
P(i, r) = P\{(i, r)| (\hat{i}, r'), r' \leq r-1\} \cdot P\{ (\hat{i}, r'), r' \leq r-1\}
\]

\[
= P\{(i, r)| (\hat{i}, r'), r' \leq r-1\} \cdot P\{(\hat{i}, r-1)| (\hat{i}, r'), r' \leq r-2\} \ldots P\{(\hat{i}, 2)| (\hat{i}, 1)\} \cdot P\{(\hat{i}, 1)\}
\]

Since,

\[
P\{(i, r)| (\hat{i}, r'), r' \leq r-1\}
\]

\[
= 1 - P\{(i, r)| (\hat{i}, r'), r' \leq r-1\} = \frac{N-r}{N-r+1}, \text{ for all } r.
\]

Therefore,

\[
P(i, r) = \frac{1}{N}
\]

This completes the proof of the lemma.

**Lemma 3.2.** For SRSWOR

\[
P\{(i, r)|i_1, i_2, \ldots, i_{r-1} \neq i\} = \frac{1}{N-r+1}
\]

for all \(i\) and \(r\) and for all the \(S(r)\) sets \((i_1, i_2, \ldots, i_{r-1} \neq i)\)

\[
\Leftrightarrow P(s_i) = \frac{1}{\binom{N}{n}} \cdot \frac{1}{n!}; \quad t = 1, 2, \ldots, \quad \left(\begin{array}{c} N \\ n \end{array}\right) \cdot n!
\]

(3.10)

where \(s_i\) consists of \(n\) distinct units of the events \(i_r, r = 1, 2, \ldots, n\) and therefore it can be denoted by \(s_i = (i_1, i_2, \ldots, i_n)\)

**Proof.** Let
\[ P\{ (i,r) | i_1, i_2, \ldots, i_{r-1} \neq i \} = \frac{1}{N - r + 1} \] for all \( i \) and \( r \) and for all the \( S(r) \) sets \((i_1, i_2, \ldots, i_{r-1} \neq i)\).

Then,
\[ P(s_t) = P(i_1) \cdot P(i_2 | i_1) \cdot \ldots \cdot P(i_n | i_{n-1}, \ldots, i_1) = \frac{1}{N \choose n} n! \]

Now, let
\[ P(s_t) = \frac{1}{N \choose n} n! \quad ; \quad t = 1, 2, \ldots, \left(\begin{array}{l} N \\ n \end{array}\right) n! \]

Then,
\[ P(i_1, i_2, \ldots, i_{r-1}) = \frac{\left(\begin{array}{c} N - r + 1 \\ n - r + 1 \end{array}\right) (n - r + 1)!}{N \choose n} n! \]

Also
\[ P(i_1, i_2, \ldots, i_r) = \frac{N - r \choose n - r} (n - r)! \]

Therefore,
\[ P(i_r | i_1, i_2, \ldots, i_{r-1}) = \frac{1}{N - r + 1} \]

That is
\[ P\{ (i,r) | i_1, i_2, \ldots, i_{r-1} \neq i \} = \frac{1}{N - r + 1} \]

This completes the proof of the lemma.

**Theorem 3.1.** The three definitions: 1(a), 2(a) and 3(a) for SRSWOR are equivalent to the classical definition of SRSWOR in (2.1) of Section 2 and therefore they are mutually equivalent.
Proof. That, Definition 1(a) is equivalent to the classical definition of SRSWOR in (2.1) of Section 2, is clear from Lemma (2.1) of that section. Also that, Definition 3(a) is equivalent to the said classical definition is clear from Lemma 3.2. As regards definition 2(a), we note \( P(i,r) = \frac{1}{N} \) if and only if \( P\{i, r| (i', r') \leq r - 1\} = 1/(N - r + 1) \) for all \( i \) and \( r \) in view of Lemma 3.1, which in turn along with the second condition of Definition 2(a) implies the said classical definition in view of Lemma 2.1. Therefore, this definition too is equivalent to the said classical definition of SRSWOR. Thus, the three alternate definitions of SRSWOR in this section are equivalent to the classical definition of SRSWOR in (2.1) of section 2 and therefore they are mutually equivalent.

In order to prove that the three definitions 1(b), 2(b) and 3(b) for SRSWR are equivalent to the classical definition of SRSWR in (2.6) of Section 2, we observe that Definition 1(b) is exactly the same as the said classical definition. The proof, that the remaining two definitions are equivalent to the said classical definition, lies in the proof of the following two lemmas.

Lemma 3.3. For SRSWR

\[
P(i, r) = \frac{1}{N}, \text{ for all } i \text{ and } r
\]

\[\Leftrightarrow P\{(i, r)|i_{r-1},\ldots,i_1\} = \frac{1}{N}, \text{ for all } i \text{ and } r\] (3.5)

Proof. In case of sampling with replacement, the n random experiments forming the compound experiment \( E \) in Sec. 2 are independent and therefore,

\[
P\{(i, r)|i_{r-1},\ldots,i_1\} = \frac{P[(i, r), i_{r-1},\ldots,i_1]}{P(i_{r-1},\ldots,i_1)} = \frac{\prod_{s=1}^{r} P(i, s)}{\prod_{s=1}^{r-1} P(i, s)} = \frac{1}{N},
\]

when \( P(i, s) = 1/N \) for all \( i \) and \( s \).

Now,

\[
P(i, r) = \sum_{i_{r-1},\ldots,i_1} [P\{(i, r)|i_{r-1},\ldots,i_1\} P\{(i_{r-1})|i_{r-2},\ldots,i_1\} \ldots P(i_2|i_1) P(i_1) ];
\]

summation being over \( N^{r-1} \) sets of \( (i_{r-1},\ldots,i_1) \)

\[
\frac{1}{N}
\]

when the converse is true

This completes the proof of the lemma.
Lemma 3.4. For SRSWR

\[ P\{ (i,r) \mid i_1, \ldots, i_{r-1} \} = \frac{1}{N} \quad \text{for all } i \text{ and } r \]

\[ \Leftrightarrow P(s_i) = \frac{1}{N^n} \quad \text{for all } t \quad (3.11) \]

**Proof**: The proof of the lemma is on the lines of that of Lemma 3.2.

Remark 3.1. It may be noted that in the literature on sampling [cf. Cochran (1977), Ch. 2, p. 18; Sukhatmes and Ashok (1984), Def. 2.1, p-24] SRSWOR is defined by

\[ P(s'_t) = \frac{1}{\binom{N}{n}}, \quad (3.12) \]

for \( t = 1, 2, \ldots, \binom{N}{n} \) giving all the unordered samples \( s'_t \) in the compound random experiment \( E \).

This definition of SRSWOR is inconsistent with Definition 3(a) of SRSWOR. However, this inconsistency does not create any problem for an unordered estimator, like sample mean, of the population mean. However, one can easily build ordered estimators of population mean such as one given by

\[ e = \frac{1}{N} \left( X_i + \frac{N-1}{n-1} \sum_{r=2}^{n} X_r \right) \quad (3.13) \]

where \( X_r \) denotes the value of the population unit drawn at the \( r \)-th draw for \( r=1, 2, \ldots, n \).

Obviously, we can not use here in such cases the above definition of SRSWOR and also Definition 1(a). However, this ordered estimator \( e \) is easily seems to be unbiased from the classical definition and Definitions 2(a) and 3(a). Such ordered estimators extensively occur in the Theory of Linear Estimation of T-Classes [cf. Bhargava , N.K. and Tikkiwal, B.D., (1978) and Bhargava, N.K. (1978) ]
Remark 3.2. The Eq. (3.7) that \( P(s'_t) = \frac{1}{\binom{N}{n}} \) for \( t = 1, 2, \ldots \) for defining SRSWOR in the literature and the conditions occurring first in the Definitions 1(a) and 2(a) of SRSWOR are necessary but not sufficient for defining SRSWOR and thus they by themselves are properties of SRSWOR.

Proof. The Eq.(3.7) in the remark follows from Definition 3(a) by noting that any particular unordered sample \( s'_t \) in Eq.(3.7) consists of some \( n! \) ordered samples. Also that, the conditions occurring under Definitions 1(a) and 2(a) respectively follow from the classical definition, is clear from Lemmas (2.1) and (3.1). Thus, Eq. (3.7) and the said two conditions are necessary for defining SRSWOR. However, they are not sufficient in view of the following example of a random sampling technique, where they hold, but the technique is not SRSWOR technique.

Example 3.1. Let us draw a sample of size 2 from a population \( \pi : (u_1, \ldots, u_6) \) of six units as follows. At the first draw, \( P(i,1) = \frac{1}{6} \) for all \( i \). At the second draw, \( P_{i}(i,2) = \frac{1}{6} \) for all \( i \).

\[
P\{(i,2) | i \neq i\} = \begin{cases} \frac{1}{5}, & \text{if } i = 1, 2, 3 \\ 0, & \text{if } i = 5 \\ \frac{2}{5}, & \text{if } i = 6 \end{cases} \tag{1}
\]

\[
P\{(4,2) | i \neq i\} = \begin{cases} \frac{1}{6}, & \text{if } i = 1, 2, 3 \\ 0, & \text{if } i = 6 \end{cases} \tag{2}
\]

\[
P\{(5,2) | i \neq i\} = \begin{cases} \frac{1}{6}, & \text{if } i = 1, 2, 3 \\ 0, & \text{if } i = 6 \end{cases} \tag{3}
\]

\[
P\{(6,2) | i \neq i\} = \begin{cases} \frac{1}{6}, & \text{if } i = 1, 2, 3 \\ 0, & \text{if } i = 6 \end{cases} \tag{4}
\]

Since, \( P\{(i,2) | i \neq i\} \) is not same for all \( i \) in this random sampling technique, as required by Eq. (2.1) of Sec. 2, it is not SRSWOR technique. However, \( P(s'_t) = \frac{1}{15} \) for all unordered samples \( s'_t \) of the type \( (u_i, u_j) \), \( i \neq j \). Thus Eq. (3.7) is satisfied. Further, \( P\{(i,r) | (i',r'), r' \leq r - 1\} = \frac{1}{5} \) and \( P\{(i,r) = 1/6 \) for all \( i \) and \( r \). Thus the required two conditions are also satisfied.

4. Justification of the Methods of Selecting Simple Random Samples
The classical urn method or its variant chit method, described in Sec. 1, for drawing random samples leads to simple random samples is clear from the way the classical definition of the two sampling techniques, SRSWOR and SRSWR, are defined in Sec. 2 by Eqs.(2.1) and (2.6). As regards the other methods selecting a simple random sample, by associating one or more random numbers to population units [cf. Murthy (1967), Cochran (1977)], as discussed in Sec. 1, admit the following theoretical framework.

Let $N$ units of the population be numbered as $u_1, u_2, \ldots, u_N$. Let the number $N$ consists of $m$ digits. Since a digit in a number can be any one of the digits $(0, 1, 2, \ldots, 9)$; such $m$-digit non-negative numbers are $10^m - 1$ in number. Let $N' = 10^m - 1$. Each method for the selection of simple random sample with the help of random number tables associates a set of particular $S(>0)$ $m$-digit numbers to each of the units in the population. In the method of associating only one random number to each population unit, $S=1$. In each method there is a set of particular $S'$ $m$-digit numbers which are not associated to any unit in the population. Thus,

$$N' = NS + S' \quad (4.1)$$

We choose a set of $m$ columns from a random number table and then go over to other sets of $m$ columns, if necessary, for the selection of $m$-digit numbers. We pick up $m$-digit numbers one by one till they result in the selection of a sample of pre-assigned size from the population of size $N$.

We refer to picking up a $m$-digit random number from the table in the above manner, as a draw. We refer to draw as a success if it results in the selection of a unit from the population and as a failure if it does not. After the first success, the unit selected should be a fresh one not selected earlier in case of a sample in which no repetition is permitted; but it need not be so in case of a sample in which repetition is permitted. This way, we need $n$ successes in order to get a sample of size $n$ in both the cases.

Since, all $m$-digit numbers in the random number table are equally likely, the probability of selecting a particular unit of the population in any particular draw is

$$\frac{S}{10^m - 1} = \frac{S}{N'} \quad (4.2)$$

and that selecting no unit at that draw is

$$\frac{S'}{10^m - 1} = \frac{S'}{N'} \quad (4.3)$$

Let

$$[r, i]$$ denote the event of getting $r$-th success with $i$-th unit of the population for $r=1, 2, \ldots, n$ and $i = 1, 2, \ldots, N$.

$$[r, i]$$; $q_r$ denote the event of getting $r$-th success with $i$-th unit of the population after $q_r$ failures for $r=1, 2, \ldots, n$ and $i = 1, 2, \ldots, N$. 
Then

\[ \Pr \{ [r, i]; q_r, |[1, i_1],...,|r-1, i_{r-1}] \} = \]

\[
= \begin{cases} 
\left( \frac{S' + (r-1)S}{N'} \right)^{q_r} \frac{S}{N'}, & \text{in case repetition of units is not permitted} \\
\left( \frac{S'}{N'} \right)^{q_r} \frac{S}{N'}, & \text{in case repetition of units is permitted}
\end{cases} 
\tag{4.4}
\]

As observed earlier, for the selection of a sample of pre-assigned size, we pick up m-digit numbers one by one. A particular m-digit number so picked up may or may not lead to the selection of unit from the population. Therefore, this process of drawing of m-digit numbers can continue indefinitely without leading to the selection of a unit. In order to ensure that this never happens, we are required to prove that this process of drawing of m-digit numbers till we get a unit from the population for the first time terminates with probability one.

In fact, we have to prove this for any r-th success for \( r = 1, 2, ..., n \). For this we state below the following theorem.

**Theorem 4.1.** The probability of getting \( r \)-th success for \( r = 1, 2, ..., n \) is one for both the cases, one not permitting the repetition of units in the sample and other permitting it.

**Proof.** **Case 1:** Repetition of units not permitted in the sample

From (4.4),

\[ \Pr \{ [r, i]; [1, i_1],...,|r-1, i_{r-1}] \} = \sum_{q_r=0}^{\infty} \left( \frac{S' + (r-1)S}{N'} \right)^{q_r} \frac{S}{N'} \]

\[ = \frac{1}{N' - r + 1}, \text{ for } r = 1, 2, ..., n \tag{4.5} \]

Therefore, for all \( r \)

\[ P \{ [r, i]; [1, i_1],...,|r-1, i_{r-1}] \}

\[ = P \{ [r, i] \mid |[1, i_1],...,|r-1, i_{r-1}] \} P \{ |[1, i_1],...,|r-1, i_{r-1}] \}

\[ = \prod_{s=2}^{r} P \{ [s, i_s] \mid |[1, i_1],...,|s-1, i_{s-1}] \} P(1, i_1) = \frac{(N-r)!}{N!} \tag{4.6} \]

where \( i_s = i \).

Now,

\[ P[r, i] = P \{ U \{ [r, i]; [1, i_1],...,|r-1, i_{r-1}] \} \]
the union being over \( \binom{N-1}{r-1}(r-1)! \) mutually exclusive sets of \((i_1, i_2, \ldots, i_{r-1}, i)\)
for different ordering of the first \((r-1)\) units.

Therefore, for all \((r,i)\)

\[
P[r,i] = \frac{1}{N} \tag{4.7}
\]

Now, the probability of \(r\)-th success is

\[
P[r] = P\{ \bigcup_{i=1}^{N} [r,i] \} = 1 \tag{4.8}
\]

**Case 2.** Repetition of the units permitted in the sample

Again, from (4.4),

\[
P\{[r,i] \mid [1,i_1], \ldots, [r-1, i_{r-1}]\} = \sum_{q_i=0}^{\infty} \left( \frac{S'}{N'} \right)^{q_i} \frac{S}{N'} = \frac{1}{N} \tag{4.9}
\]

for \(r = 1, 2, \ldots, n\).

Therefore, for all \(r\)

\[
P\{[r,i][1,i_1] \ldots [r-1, i_{r-1}]\} = \prod_{s=2}^{r} \{P [s,i_s][1,i_1], \ldots, [s-1, i_{s-1}]\} P[1,i_1] = \frac{1}{N^r} \tag{4.10}
\]

Now,

\[
P[r,i] = P\{U[r,i][1,i_1] \ldots [r-1, i_{r-1}]\}
\]

the union being over \(N^{r-1}.N^{r-1} = N^{r-1}\) mutually exclusive sets of \((i_1, i_2, \ldots, i_{r-1}, i)\)
for different ordering of the first \((r-1)\) units.

Therefore,

\[
P[r,i] = \frac{1}{N} \tag{4.11}
\]

which in turn gives, as before

\[
P[r] = 1
\]

This completes the proof.

**Remark 4.1.** We note that the event \(\{[r,i][1,i_1] \ldots [r-1, i_{r-1}]\}\) corresponds to the occurrence of the sample \(s_i = (i_1, i_2, \ldots, i_n)\) for \(r = n\) and \(i = i_n\). Therefore, from (4.6).


\[ P(s_t) = \frac{1}{\binom{N}{n} n!}, \text{ for all } t \]  

(4.12)

in case the repetition of units is not permitted, that is in case of sampling without replacement.

Also from (4.10), we have

\[ P(s_t) = \frac{1}{N^n}, \text{ for all } t \]  

(4.13)

in case the repetition of units is permitted in the sample, that is, in case of sampling with replacement.

**Remark 4.2.** When the repetition of the units is not permitted in the selection of a sample through random numbers we have relation (4.12) corresponding to Definition 3(a) for SRSWOR given in Section 3. Therefore, from this relation, we conclude that these methods of selection lead to samples through SRSWOR. But, when the repetition of the units is permitted in the selection of a sample through random numbers; we have relation (4.13) corresponding to Definition 3(b) for SRSWR given also in Section 3. Therefore, from this relation, we conclude that the above methods of selection lead to the samples also through SRSWR. Thus, we justify that the methods of selection considered in this section lead to simple random samples.

**ACKNOWLEDGMENT**

The author is grateful to Prof. B.D. Tikkiwal for enlightening discussions at the time of preparing the manuscript of the paper. The author is also grateful to the referees for making valuable observations; which helped in improving the presentation of the paper.

**REFERENCES**


THE USE OF A KNOWN COEFFICIENT OF VARIATION IN THE ESTIMATION OF MEAN OF A NORMAL DISTRIBUTION FROM DOUBLE SAMPLES

L. N. Upadhyaya\textsuperscript{1} and S. R. Srivastava\textsuperscript{2}

ABSTRACT

In this paper we have proposed an estimator of the population mean $\mu$ from double samples when the coefficient of variation of the normal distribution is known and we have a guessed value of the population mean $\mu$. The properties of the proposed estimator has been discussed. Numerical values of relative efficiencies have been obtained and recommendations regarding its use have been made.

Keywords: Population mean, coefficient of variation, mean-squared error, relative efficiency.

1. Introduction

When there is no apriori knowledge available for the population mean, it is well known that the sample mean based on all sample observations provides the best estimate. However, in some practical problems there is a guess of the population mean available either from the past experience or due to acquaintance of the experimenter with the nature of data he has to handle. Katti (1962) gave a procedure for estimating the mean of a normal population when its variance is known under the conditions that the experimenter has a sample available to him and could take a second sample, if needed. His procedure consists in the construction of a region $R_0 \subset \mathbb{R}$\textsuperscript{1} (the real line) based on the guessed value $\mu_0$ of the population mean, sample sizes $n_1$, $n_2$ of the two samples and the population standard deviation $\sigma$. It is now assumed that the coefficient of variation $C (= \sigma /\mu)$ of the population rather than the standard deviation $\sigma$ is known; then it is possible to propose an estimator of the mean following the procedure given by

\textsuperscript{1} Professor, Department of Applied Mathematics, Indian School of Mines, Dhanbad – 826 004. India.

\textsuperscript{2} (Retired) Professor of Statistics, Banaras Hindu University, Varanasi – 221 005. India.
Katti (1962). The estimator has been described in section 2 and its properties have been discussed in section 4.

2. Estimation Procedure

When C is known, various estimators of the population mean $\mu$ have been proposed by Khan (1968), Govindrajulu and Sahai (1972), and Sen (1978). One of the unbiased estimators for $\mu$ considered by them is

$$\hat{\mu} = ks$$

based on n sample observations where

$$k = \sqrt{\frac{n-1}{2}} \sqrt{\frac{n-1}{2}}$$

and

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Let $x_{1j}$, $j = 1, 2, \ldots, n_1$, be the first sample and $x_{2j}$, $j = 1, 2, \ldots, n_2$, the second sample observations and they are all independent and identically distributed with mean $\mu$ and variance $C^2 \mu^2$. Let $\hat{\mu}_1 = k_1 s_1$ and $\hat{\mu}_2 = k_2 s_2$ be the two unbiased estimators of $\mu$ where

$$k_i = \frac{N_i}{C}$$

$$N_i = \frac{\sqrt{n_i-1}}{2} \frac{\sqrt{n-1}}{2}$$

$$s_i^2 = \frac{1}{(n_i-1)} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

(i = 1, 2)

However, the minimum variance unbiased estimator of $\mu$, using a linear combination of $\hat{\mu}_1$ and $\hat{\mu}_2$, is given by

$$\hat{\mu} = g_1 s_1 + g_2 s_2,$$
where
\[
g_1 = \frac{N_2^2 - 1}{N_1^2 + N_2^2 - 2} k_1,
\]
and
\[
g_2 = \frac{N_1^2 - 1}{N_1^2 + N_2^2 - 2} k_2.
\]

Assuming the guessed value of the mean to be \( \mu_o \), we propose an estimator for \( \mu \) as follows:
\[
\hat{\mu}_p = \begin{cases} 
  k_1 s_1, & \text{if } k_1 s_1 \in R \\
  g_1 s_1 + g_2 s_2, & \text{if } k_1 s_1 \not\in R
\end{cases}
\]
where \( R \subset R^1 \) is the region specified in section 3.

3. The Region R

If \( \mu_o \) is the true value of the mean \( \mu \), the expected mean-squared error of \( \hat{\mu}_p \) is given by
\[
E. M. S. \left( \hat{\mu}_p \mid \mu_o \right) = \int_{s_1=0}^{\infty} \int_{s_2=0}^{\infty} (\hat{\mu}_p - \mu_o)^2 p_1(s_1)p_2(s_2) \, ds_2 \, ds_1
\]
\[
= \int_{s_1 \in R} \int_{s_2=0}^{\infty} (k_1 s_1 - \mu_o)^2 p_1(s_1)p_2(s_2) \, ds_2 \, ds_1
\]
\[
+ \int_{s_1 \in \overline{R}} \int_{s_2=0}^{\infty} (g_1 s_1 + g_2 s_2 - \mu_o)^2 p_1(s_1)p_2(s_2) \, ds_2 \, ds_1
\]
\[
= \int_{s_1 \in R} \int_{s_2 = 0}^{\infty} (k_1 s_1 - \mu_0)^2 p_1(s_1) d(s_1) + \frac{N^2 - 1}{(N_1^2 + N_2^2 - 2)^2}.
\]

We have to choose \( R \) such that (3.1) is minimum. This results in

\[
\text{R: } \left[ 1 - \frac{\sqrt{N_{12}}}{k_1} \right] \mu_0 \leq s_1 \left[ 1 + \frac{\sqrt{N_{12}}}{k_1} \right] \mu_o,
\]

where

\[
N_{12} = \frac{(N_1^2 - 1)(N_2^2 - 1)}{(N_1^2 + 2N_2^2 - 3)}
\]

Thus, our estimator \( \hat{\mu}_p \) for the population mean \( \mu \) is

\[
\hat{\mu}_p = \begin{cases} 
    k_1 s_1, & \text{if } (1 - \sqrt{N_{12}}) \mu_0 \leq k_1 s_1 \leq (1 + \sqrt{N_{12}}) \mu_o \\
    g_1 s_1 + g_2 s_2, & \text{otherwise}
\end{cases}
\]

4. Properties of the Estimator \( \hat{\mu}_p \)

If \( \mu_o \) is the true value of the mean, the expected mean squared error of the estimator \( \hat{\mu}_p \) is given by

\[
E. M. S. (\hat{\mu}_p | \mu_o) = \int_{s_1 = B_1}^{B_2} \int_{s_2 = 0}^{\infty} (k_1 s_1 - \mu_0)^2 p_1(s_1) p_2(s_2) d(s_2) d(s_1)
\]
\[ + \int_{s_1=B_2}^{\infty} \int_{s_2=0}^{s_1} (g_1 s_1 + g_2 s_2 - \mu_0)^2 p_1(s_1) p_2(s_2) d(s_2) d(s_1) \quad (4.1) \]

where

\[
B_1 = \left(1 - \sqrt{\frac{N_2}{12}}\right) \mu_0 \quad \text{and} \quad B_2 = \left(1 + \sqrt{\frac{N_2}{12}}\right) \mu_0
\]

If we integrate (4.1) and simplify, we get

\[ \text{E. M. S.} (\hat{\mu}_P | \mu_0) = \frac{N_1^2 - 1}{(N_0^2 + N_0^2 - 1)^2} T_1 \mu_0^2, \]

where

\[
T_1 = A_1 + A_2 D_{-1} + A_3 (2D_0 - N_1^2 D_1),
\]

\[
A_1 = (N_2^2 - 1)(N_1^2 + N_2^2 - 2),
\]

\[
A_2 = N_1^2 N_2^2 - 3N_2^2 - 2N_1^2 + 4,
\]

\[
A_3 = N_1^2 + 2N_2^2 - 3,
\]

\[
D_j = \frac{P\left(\frac{n_1 + j}{2}, t_a\right) - P\left(\frac{n_1 + j}{2}, t_b\right)}{(j = -1, 0, 1)}
\]

\[
P(N, X) = \frac{1}{\sqrt{N}} \int_0^x e^{-\frac{1}{2} \left(\frac{N - 1}{2}\right) t^2} dt,
\]

\[
t_a = \frac{(n_1 - 1)(1 - \sqrt{\frac{N_2}{12}})^2}{N_1^2},
\]

and
The relative efficiency of $\hat{\mu}_p$ with respect to the MVU estimator $\hat{\mu}_3$ is given by

$$E. (\hat{\mu}_p | \mu_0) = \frac{(N_2^2 - 1)(N_1^2 + N_2^2 - 2)}{T_1} \mu^2,$$

where

$$T_2 = A_1 + A_2 D' - A_3 (2D'_0 - N_1^2 D'_1)$$

$$D'_j = P\left(\frac{n_1 + j}{2}, t'_a\right) - P\left(\frac{n_1 + j}{2}, t'_b\right) \quad (j = 1, 0, 1)$$

$$t'_a = \delta t_a, \quad t'_b = \delta t_b, \quad \delta = \left(\frac{\mu_0}{\mu}\right)^2.$$

The relative efficiency of $\hat{\mu}_p$ with respect to $\hat{\mu}_3$ is given by

$$E (\hat{\mu}_p | \mu) = \frac{(N_2^2 - 1)(N_1^2 + N_2^2 - 2)}{T_2}.$$

5. Discussion of the Results:
Table 1 gives the values of the relative efficiencies of \( \hat{\mu}_p \) with respect to \( \hat{\mu}_3 \), \( E(\hat{\mu}_p \mid \mu_0) \) given by (4.2) for different values of \( n_1 \) and \( n_2 \). From this table, it follows that the proposed estimator \( \hat{\mu}_p \) is always more efficient than \( \hat{\mu}_3 \). For fixed \( n_1 \), the relative efficiency increases as \( n_2 \) increases; whereas if \( n_2 \) is fixed, the relative efficiency decreases with the increase in \( n_1 \). For \( n_1 = n_2 \), the value of the relative efficiency remains almost the same. Thus, when \( \mu_0 \) is the value of the true mean, it is preferable to take \( n_2 \) greater than \( n_1 \).

In table 2, we have tabulated the values of the relative efficiencies \( E(\hat{\mu}_p \mid \mu) \), given by (4.3), for different values of \( n_1, n_2 (> n_1) \) and \( \delta \). Here the proposed estimator \( \hat{\mu}_p \) turns out to be more efficient than \( \hat{\mu}_3 \) if we take \( \delta \) near 1.0 (0.75 < \( \delta \) < 1.25). For other values of \( \delta \) outside the said range the preference goes to \( \hat{\mu}_3 \).

**Table 1:** Values of \( E(\hat{\mu}_p \mid \mu_0) \)

<table>
<thead>
<tr>
<th>( n_2 )</th>
<th>( n_1 = 5 )</th>
<th>( n_1 = 10 )</th>
<th>( n_1 = 15 )</th>
<th>( n_1 = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.2065</td>
<td>1.1087</td>
<td>1.0783</td>
<td>1.0608</td>
</tr>
<tr>
<td>10</td>
<td>1.2053</td>
<td>1.1724</td>
<td>1.1379</td>
<td>1.1150</td>
</tr>
<tr>
<td>15</td>
<td>1.2022</td>
<td>1.1982</td>
<td>1.1728</td>
<td>1.1509</td>
</tr>
<tr>
<td>20</td>
<td>1.1916</td>
<td>1.2080</td>
<td>1.1967</td>
<td>1.1724</td>
</tr>
</tbody>
</table>

**Table 2:** Values of \( E(\hat{\mu}_p \mid \mu) \)

<table>
<thead>
<tr>
<th>( \xi )</th>
<th>( n_1 = 5 )</th>
<th>( n_1 = 10 )</th>
<th>( n_1 = 15 )</th>
<th>( n_1 = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.7011</td>
<td>0.6377</td>
<td>0.5964</td>
<td>0.8486</td>
</tr>
<tr>
<td>0.50</td>
<td>0.7807</td>
<td>0.7118</td>
<td>0.6584</td>
<td>0.7229</td>
</tr>
<tr>
<td>0.75</td>
<td>1.0444</td>
<td>1.0025</td>
<td>0.9341</td>
<td>0.9533</td>
</tr>
<tr>
<td>1.00</td>
<td>1.2053</td>
<td>1.2022</td>
<td>1.1916</td>
<td>1.1982</td>
</tr>
<tr>
<td>1.25</td>
<td>1.0691</td>
<td>1.5555</td>
<td>0.9946</td>
<td>0.9869</td>
</tr>
<tr>
<td>1.50</td>
<td>0.8493</td>
<td>0.7732</td>
<td>0.7179</td>
<td>0.7419</td>
</tr>
<tr>
<td>2.00</td>
<td>0.5881</td>
<td>0.5036</td>
<td>0.4493</td>
<td>0.5927</td>
</tr>
<tr>
<td>3.00</td>
<td>0.4970</td>
<td>0.4241</td>
<td>0.3780</td>
<td>0.7719</td>
</tr>
</tbody>
</table>
REFERENCES


KATTI, S. K. (1972), Use of some a priori knowledge in the estimation of means from double samples. Biometrics, 18, pp. 139-47.

ESTIMATION OF POPULATION MEAN ON THE BASIS
OF NON-SIMPLE SAMPLE WHEN NON-RESPONSE
ERROR IS PRESENT

Janusz Wywiał

ABSTRACT

The entire sampling design and sampling scheme are determined by the
appropriate set of the probabilities of inclusion of the first and second order.
The estimator of a population mean is considered in the case when a non-
response error is present. At the first stage a sample is selected from the
population. The estimator of a population mean determined on the basis of the
data observed in units that respond, is determined. In the second phase, a
simple sample is drawn from the set of units that do not respond. Next, the
second estimator of the mean from the second phase sample is determined.
The linear combination of both these statistics is the estimator of the
population mean. The coefficients of this linear combination should be close
to the fractions of the strata. Some procedures of using auxiliary variables
observed in the whole population are suggested to determine these
coefficients. The expected value and the mean square error of the estimator are
derived.

Key words: non-response error, call-back method, logistic super-population
model, conditional sampling design, sub-sampling of nonrespondents.

1. Estimator

Let \( U = \{1, 2, \ldots, N\} \) be a finite population. It is divided into two disjoint
subpopulations.

The partition \( \{U_A, U_B\} \) of the population is determined by a response
distribution denoted by “RE”. The sizes of \( U_A \) and \( U_B \) are denoted by \( N_A \) and \( N_B \),
respectively. Moreover, let \( U_A \neq \emptyset \) and \( U_A \cap U_B = U \). The observation \( y_k \) of a
variable under study is attached to the \( k \)-th population element. Let \( S \) be a sample
of a fixed size drawn from the population \( U \) according to a sampling design \( P_1(s) \).

1 Department of Statistics; University of Economics, Bogucicka 14, 40-226 Katowice, Poland
e-mail: wywial@ae.katowice.pl
Its inclusion probabilities of the first and second sizes are denoted by: \( \pi_k, \pi_{kt}, k \neq t = 1, \ldots, N \). Let \( s_A, s_B \) be such a partition of the sample \( s \) that \( s_A \subseteq U_A, s_B \subseteq U_B, s = s_A \cup s_B \). So, \( s_A \cap s_B = \emptyset \). The sizes of \( s_A \) and \( s_B \) are denoted by \( n_A \) and \( n_B \), respectively. We assume that \( U_A \) consists of respondent elements and \( U_B \) of non-respondent elements. If \( s_B \neq \emptyset \), we assume that the sample \( s_2 \) is selected on the basis of the conditional sampling design \( \mathbb{P}_2(s_2 | s_B) \) with the inclusion probabilities \( \pi_{k/s_B}, \pi_{kt/s_B} \). We shall use the following notation introduced by Särndal, Swenson and Wretman (1992):

\[
\pi_{ek} = \pi_k \pi_{k/s_B} \quad \text{if} \quad k, t \in s_B
\]

\[
\pi_{ekt} = \begin{cases} 
\pi_{kt} \pi_{k/s_B} & \text{if} \quad k \in s_B, t \in s_A \\
\pi_{kt} \pi_{t/s_B} & \text{if} \quad k \in s_A, t \in s_B \\
\pi_{kt} & \text{if} \quad k, t \in s_A 
\end{cases}
\]

We assume that all the values of the variable are observed in the sample \( s_2 \).

Let:

\[
\bar{y}_{s_A} = \frac{\sum_{k \in s_A} \hat{y}_k}{\sum_{k \in s_A} \pi_k} \quad ; \quad \bar{y}_{s_B} = \frac{\sum_{k \in s_B} \hat{y}_k}{\sum_{k \in s_B} \pi_k}
\]

\[
\bar{y}_{s_2} = \frac{\sum_{k \in s_2} \hat{y}_k}{\sum_{k \in s_2} \pi_k}
\]

where:

\[
\hat{y}_k = \frac{y_k}{\pi_k}, \quad \hat{y}_{kt} = \frac{y_k}{\pi_{kt}} \quad \text{and} \quad \hat{y}_k/s_B = \frac{y_k}{\pi_{k/s_B}}
\]

Let us consider the following estimator of the population mean \( \bar{y} = \frac{1}{N} \sum_{k=1}^{N} \hat{y}_k \) :

\[
\bar{y}_{s*} = \alpha \bar{y}_{s_A} + (1 - \alpha) \bar{y}_{s_2}, \quad s_* = \{s, s_2\}
\]
where: \(0 < \alpha \leq 1\) can be dependent on the samples (s, \(s_2\)).

In order to derive the basic parameters of the statistic \(\bar{y}_{s_2}\), we consider the following Taylor expansion of the sample means (see Särndal et all, 1992, p. 178):

\[
\bar{y}_{s_2} \approx \bar{y}_{U_A} + \frac{1}{N_A} \sum_{k \in U_A} y_k - \bar{y}_{U_A} \pi_k
\]

(7)

or:

\[
\bar{y}_{s_2} \approx \bar{y}_{U_A} + \frac{1}{N_B} \sum_{k \in U_B} y_k - \bar{y}_{B} \pi_k + \frac{1}{N_B} \sum_{k \in U_B} y_k - \bar{y}_{n} \pi_k
\]

(8)

where:

\[
\bar{y}_{A} = \frac{1}{N_A} \sum_{k \in U_A} y_k, \quad \bar{y}_{B} = \frac{1}{N_B} \sum_{k \in U_B} y_k
\]

\[
N_A = \sum_{k \in U_A} \pi_k, \quad N_B = \sum_{k \in U_B} \pi_k
\]

Now, we have (see Särndal et all, 1992, p. 182 – 184):

\[
E_{s_2}(\bar{y}_{s_2}) = \bar{y}_{s_2}, \quad E_{s_2}(\bar{y}_{s_2}) = \bar{y}_{s_2}
\]

(10)

\[
E_{RD}(\bar{y}_{s_2}) = \bar{y}_{s_2}, \quad E_{RD}(\bar{y}_{s_2}) = \bar{y}_{s_2}
\]

(11)

\[
E_s(\bar{y}_{s_2}) = \bar{y}_{U_A}, \quad E_s(\bar{y}_{s_2}) = \bar{y}_{U_A}
\]

(12)

So:

\[
E_s E_{RD} E_{s_2}(\bar{y}_{s_2}) = \bar{y}_{U_A}
\]

(13)

Let us assume that \(\alpha\) is fixed.

\[
D^2(\bar{y}_{s_2}) = \alpha^2 D^2(\bar{y}_{s_2}) + 2\alpha(1 - \alpha) \text{Cov}(\bar{y}_{s_2}, \bar{y}_{s_2}) + (1 - \alpha)^2 D^2(\bar{y}_{s_2})
\]

(14)

\[
D^2(\bar{y}_{s_2}) = \frac{1}{N_A^2} \sum_{k_1 \in U_A, k_2 \in U_A} \Delta_{k_1 k_2} \left( \frac{y_k - \bar{y}_{U_A}}{\pi_k} \right) \left( \frac{y_t - \bar{y}_{U_A}}{\pi_t} \right)
\]

(15)
where:

\[ \tilde{\Lambda}_{kt} = \pi_{kt} - \pi_k \pi_t \]

\[ \text{Cov}(\bar{y}_{sa}, \bar{y}_{s}) = \text{Cov}(E_{RD}E_{s_{a}}(\bar{y}_{sa}), E_{RD}E_{s_{2}}(\bar{y}_{s})) + \right. \]

\[ + E_{S}(\text{Cov}_{RD_{s_{2}}}(\bar{y}_{sa}, \bar{y}_{s})) = \]

\[ = \text{Cov}(E_{s}(\bar{y}_{sa}, \bar{y}_{s})) + E_{S}(\text{Cov}_{RD}(\bar{y}_{sa}, \bar{y}_{s})) + E_{S}E_{RD}(\bar{y}_{sa} \text{Cov}_{s_{2}}(1, \bar{y}_{s})) = \]

\[ = \text{Cov}(\bar{y}_{sa}, \bar{y}_{s}) + E_{S}(E_{RD}(\bar{y}_{sa} - \bar{y}_{s})(\bar{y}_{s} - \bar{y}_{s})) + 0 = \]

\[ \text{Cov}(\bar{y}_{sa}, \bar{y}_{s}) \]

(16)

\[ D^2(\bar{y}_{sa}) = D^2_{S}(E_{RD}E_{s_{a}}(\bar{y}_{s})) + E_{S}(D^2_{RD_{s_{2}}}(\bar{y}_{s})) = \]

\[ = D^2_{S}(\bar{y}_{s}) + E_{S}(D^2_{RD}(\bar{y}_{s})) + E_{S}(E_{RD}(D^2_{s_{2}}(\bar{y}_{s}))) = \]

\[ = D^2_{S}(\bar{y}_{s}) + E_{S}(E_{RD}(D^2_{s_{2}}(\bar{y}_{s}))) = \]

\[ = D^2_{S}(\bar{y}_{s}) + E_{S}(E_{RD}(D^2_{s_{2}}(\bar{y}_{s}))) \]

(17)

The expressions (7) – (9), (16) and the results of Särndal et al. (1992), p. 171, 178, 184 lead to the following

\[ \text{Cov}_{s}(\bar{y}_{sa}, \bar{y}_{s}) = E_{S}(\bar{y}_{sa} - E_{S}(\bar{y}_{sa}))(\bar{y}_{s} - E_{S}(\bar{y}_{s})) \approx \]

\[ \approx E_{S}\left( \frac{1}{N_A} \sum_{k \in s_{a}} y_k - \bar{y}_{U_{a}} \right)\left( \frac{1}{N_B} \sum_{k \in s_{a}} y_k - \bar{y}_{U_{a}} \right) = \]

\[ = \frac{1}{N_A N_B} \sum_{k \in U_{a}, t \in U_{a}} \sum_{t \in U_{a}} \Lambda_{kt} y_{k - \bar{y}_{U_{a}}} y_{t - \bar{y}_{U_{a}}} \]

(18)

\[ D^2(\bar{y}_{sa}) = \frac{1}{N_B} \sum_{k \in U_{a}, t \in U_{a}} \sum_{t \in U_{a}} \Lambda_{kt} y_{k - \bar{y}_{U_{a}}} y_{t - \bar{y}_{U_{a}}} \]

(19)
The expression (8) leads to the following one:

\[
D_{s_2}^2(\bar{y}_{s_2}) = E_{s_2} \left( \frac{1}{n_{B}} \sum_{k \in s_2} \frac{y_k - \bar{y}_{s_2}}{\pi_k} \right)^2 = E_{s_2} \left( \frac{1}{n_{B}} \sum_{k \in s_2} \frac{y_k - \bar{y}_{s_2}}{\pi_k} \right)^2 =
\]

\[
= \frac{1}{n_{B}} E_{s_2} \left( \sum_{k \in s_2} \frac{y_k - \bar{y}_{s_2}}{\pi_k} \right)^2 = \frac{1}{n_{B}} E_{s_2} \left( \sum_{k \in s_2} \frac{y_k - \bar{y}_{s_2} a_{k/s_2}}{\pi_k \pi_k/s_2} \right)^2
\]

where: \( a_{k/s_2} = 1 \) if \( k \in s_2 \), \( a_{k/s_2} = 0 \) if \( k \neq s_2 \), \( n_B = \sum_{k \in s_2} \frac{1}{\pi_k} \)

Hence:

\[
D_{s_2}^2(\bar{y}_{s_2}) = \frac{1}{n_B} \sum_{k \in s_1} \sum_{t \in s_2} \frac{y_k - \bar{y}_{s_2}}{\pi_k} \frac{y_t - \bar{y}_{s_2}}{\pi_t} D_{s_2}^2(a_{k/s_2}, a_{t/s_2}) =
\]

\[
= \frac{1}{n_B} \sum_{k \in s_1} \sum_{t \in s_2} \frac{y_k - \bar{y}_{s_2}}{\pi_k} \frac{y_t - \bar{y}_{s_2}}{\pi_t} \Delta_{k,t/s_2}
\]

where: \( \Delta_{k,t/s_2} = \pi_{k,t/s_2} - \pi_{k/s_2} \pi_{t/s_2} \)

The results (14), (15), (18), (19) and (20) lead to the following expression:

\[
D^2(\bar{y}_{s_1}) \approx \frac{\alpha^2}{N_A} \sum_{k \in s_1} \sum_{t \in s_2} \Delta_{k,t} \frac{y_k - \bar{y}_{A}}{\pi_k} \frac{y_t - \bar{y}_{A}}{\pi_t} +
\]

\[
+ \frac{2\alpha(1-\alpha)}{N_A N_B} \sum_{k \in s_1} \sum_{t \in s_2} \Delta_{k} \frac{y_k - \bar{y}_{A}}{\pi_k} \frac{y_t - \bar{y}_{B}}{\pi_t} +
\]

\[
+ \frac{(1-\alpha)^2}{N_B^2} \sum_{k \in s_1} \sum_{t \in s_2} \Delta_{k} \frac{y_k - \bar{y}_{B}}{\pi_k} \frac{y_t - \bar{y}_{B}}{\pi_t} +
\]

\[
+ (1-\alpha)^2 E_{s_2} E_{RD} \left( \frac{1}{n_B} \sum_{k \in s_1 \cap s_2} \frac{y_k - \bar{y}_{s_2}}{\pi_k} \frac{y_t - \bar{y}_{s_2}}{\pi_t} \right)
\]
The unbiased estimator of the above expression is as follows (see Särndal et al., 1992, p. 348).

\[
\hat{D}_{s_{i}}(\bar{y}_{s_{i}}) = \frac{\alpha^{2}}{n_{A}} \sum_{k \in A} \sum_{t \in A} \Delta_{kt} \frac{y_{k} - \bar{y}_{s_{A}}}{\pi_{k}} \frac{y_{t} - \bar{y}_{s_{A}}}{\pi_{t}} + \\
+ \frac{2\alpha(1 - \alpha)}{n_{A}^{2} n_{B}^{2}} \sum_{k \in A} \sum_{t \in A} \Delta_{kt} \frac{y_{k} - \bar{y}_{s_{A}}}{\pi_{k}} \frac{y_{t} - \bar{y}_{s_{A}}}{\pi_{t}} + \\
+ \frac{(1 - \alpha)^{2}}{n_{A}^{2}} \sum_{k \in A} \sum_{t \in A} \frac{y_{k} - \bar{y}_{s_{A}}}{\pi_{k}} \frac{y_{t} - \bar{y}_{s_{A}}}{\pi_{t}} \frac{\Delta_{kt}}{\pi_{kt}} + \\
+ \frac{(1 - \alpha)^{2}}{n_{B}^{2}} \sum_{k \in A} \sum_{t \in A} \frac{y_{k} - \bar{y}_{s_{A}}}{\pi_{k}} \frac{y_{t} - \bar{y}_{s_{A}}}{\pi_{t}} \frac{\Delta_{kt/\pi_{B}}}{\pi_{kt/\pi_{B}}}
\]

where:

\[
n_{A} = \sum_{k \in A} \frac{1}{\pi_{k}}, \quad n_{B} = \sum_{k \in B} \frac{1}{\pi_{k}}, \quad \Delta_{kt} = \frac{\Delta_{kt}}{\pi_{kt}}
\]

2. Assessing the parameter \(\alpha\)

Let us note that the parameter \(\alpha\) can be predicted by means of the logistic regression function when all observations of a multidimensional auxiliary variable are available. Let \(Z_{i} = 0\) if an \(i\)-th element of the population is non-respondent and let \(Z_{i} = 1\) when it is respondent. An \(i\)-th observation of a \(q\)-dimensional auxiliary variable is denoted by \(x_{i} = [x_{i1} \ldots x_{iq}]\). The logistic superpopulation model is as follows:

\[
P(Z_{i} = 1) = E(Z_{i}) = \frac{1}{1 + \exp\{x_{i}\beta^{T}\}}, \quad i = 1, ..., N
\]

where: \(\beta^{T} = [\beta_{1} \ldots \beta_{q}]\) is the vector of the parameters. The problem of the estimation of the vector \(\beta\) was considered e.g. by Cassel, Särndall and Wretman (1983) or Kleinbaum, Kupper, Lechtonen, Veijanen (1998) and Morgenstern (1982). Let \(\hat{\beta}_{s}\) be the estimator of \(\beta\).

When the values of the auxiliary variables are observed in a population, the parameter \(\alpha\) can be estimated by means of the statistic:

\[
\hat{\alpha}_{s} = \sum_{i=1}^{N} \hat{Z}_{si}
\]
where:

\[
\hat{Z}_{si} = \frac{1}{1 + \exp\left( x_i \beta \right)}
\]  \hspace{1cm} (25)

The statistical packages WesVar or Sudaan can be used to determine the value of the statistic \( \hat{\alpha} \) and its approximate mean-square error.

The problem of determining the variance of the statistic \( \hat{\bar{y}}_{\alpha} \), in the case when \( \hat{\alpha} \) is substituted for \( \alpha \), becomes more complicated. In order to do it the method of bootstrap or “jack-knife” can be applied.

The parameter \( \alpha \) can be assessed by means of appropriate classification methods, too.

The simulation study of the accuracy of this estimation method will be developed in a separate paper.

Acknowledgement

*The research was supported by the grant number 1 H02B 008 16 from the State Committee for Scientific Research (KBN).*

REFERENCES


SOME ASPECTS OF STATISTICAL TRAINING IN BANKING IN UKRAINE

S. Gerasymenko¹ and N. Golovach²

ABSTRACT

This article follows a series of publications about teaching applied statistics in Ukraine (SiT, June 1998). As it is known, the bank system both produces and consumes, and demands for direction the large volume of information. Its collection, analysis and use of the results of the analysis for effective management of bank activities requires statistical knowledge. Bank Statistics, a new applied statistics course, has been taught at bank schools in Ukraine for 5 years. Since 2001 Bank Statistics will be taught as a part of special programs in all universities.

Introduction

Changes in social system always demand a new mechanism of economic development and the review and change of the functions of the existing mechanisms.

In transition to market economy the new money-credit system became one of such mechanisms. One of its main components is the bank system, which:

- in the wide meaning of the word – is the aggregate of the relations of accumulation and distribution of money, the movement of debt capital and payments which are made through the bank;
- in the narrow meaning of the word – is the aggregate of the institutions which perform the above mentioned functions.

In Ukraine the need to build a viable and effectively working bank system was paramount because the realisation of economic reforms, tax, money-credit and currency policy depend to great extent upon it.

The bank system belongs to big, complex and open systems. In connection with this it produces and consumes a large volume of information. Completeness, truthfulness of information, adequate methods of its analysis are the guarantee of

¹ Kiev National University of Economics, 54/1, pr. Pobedy, 03057, Kiev, UKRAINE.
² Kiev State Economics University of Trade, 19, ul. Kioto, 02156, Kiev, UKRAINE.
making the effective managerial decisions in bank activities. In connection with this bank employees must be able to organise the gathering of information necessary for making decisions, to analyse it, and most importantly estimate risks. These actions require the ability to use the statistical methods of accounting, analysis, modelling and forecasting.

The specialists-bankers did not receive this training before. First of all because of the lack of the program of training *Bank Statistics* and the absence of the professors who could teach it.

In connection with this during the reform of the higher education in Ukraine the necessity of creating and teaching the special statistical discipline, the knowledge of which would allow the bank employees to use the above mentioned statistical methods appeared. The program of the course *Bank Statistics* was elaborated, and it has been taught as part of some bank schools and in special courses for bank employees since 1996. Bankers-practical workers eagerly took the opportunity to acquire the knowledge necessary for improving their work. The problem was to include the course *Bank Statistics* in the curriculum for training the specialists-bankers in the state institutes and universities of Ukraine. The departments of banking of these higher educational establishments, which elaborate the curriculum for training future bankers, refused to include *Bank Statistics* in it. The argument for their decision was that the terms and methods proposed by *Bank Statistics* are known to the students as they learn *Theory of Statistics, Theory of Probability, Banking, Lending* and some other subjects. But it was not taken into consideration that at the same time the economists learn *Economic Statistics*, sociologists learn *Social Statistics*, medical students learn *Medical Statistics* and so on.

But more and more bank employees learn *Bank Statistics* at the special courses and bank schools. As the result, the program of the course and the work of the professors of *Bank Statistics* – the leading Ukrainian professors of statistics – received a very high appraisal from bankers, and the directors of banks began to require from the graduates the knowledge of *Bank Statistics*. In 1998 *The National Bank of Ukraine* demanded from *The Ministry of Education* to include *Bank Statistics* as an obligatory subject in the curriculum of training bankers at master's level. Such decision was made and from 2001 the teaching of *Bank Statistics* will start in all higher educational establishments of Ukraine which prepare future bank employees.

*Bank Statistics* (in other words *Statistics of Bank Activities*) teaches the mass phenomena and processes in bank system and is the part of *Social-Economic Statistics*. 
Its aim is:
- calculating the financial indices necessary for directing bank activities,
- elaborating the statistical methods of their analysis for revealing the regularity of bank functioning,
- well-grounded forecasting of activities with taking into account the influence of factors on the basis of the corresponding statistical models.

The Subject of statistical study of bank activities is determined by the functions of banks. These are the processes of:
- creating bank capital,
- functioning of debt capital,
- receiving profits,
- operating the bank risks,
- mastering the market of bank services
and also the interrelation with customers, investors, and share-holders.

But Bank Statistics characterises not only the quantitative, but also qualitative sides of bank activities:

the quantitative side – is the aggregate of numerical information about the number of banks and number of their customers, the value of their assets, liabilities, profits, the level of interest rate, the value and terms of issued loans, and so on;

the qualitative side – is the mechanism of developing of a particular phenomenon, their regularities and tendencies in the form of generalising characteristics, e.g. – the state of the bank system, the rating of the bank, the level of reliability of the bank customers, the level of risk of lending, the state of assets and liabilities, and so on.

So, using the corresponding methodological instructions and the necessary information, it is possible to make statistical analysis and forecasting of bank activities, and on the basis of the conclusion attached to elaborate the strategy of a given bank or the bank system as a whole.

The structure of the course Bank Statistics includes the following parts:

1. Subject, method, tasks and system of indices of bank activities.
   The traditional introduction to any course where it is grounded the necessity of learning Bank Statistics. The characteristic of the existing sources of information, functions of bank activities and statistic methods of analysis of its indices is given.

2. Statistical of analysis of the balance sheet.
   The methods of modelling the correlation between active and passive operations of a bank, analysis of the quality of assets, estimation of the level of reliability, and on their basis – the rating of the bank are considered.
3. **Statistics of the risks of bank activities.**
   It contains the information about the use of statistical methods for obtaining the quantitative meaning of risks for different spheres of bank activities.

4. **Statistical analysis of credit activities of the bank.**
   It’s one of the most important parts of *Bank Statistics*, in which it is considered:
   - the use of index method for examining the turnover of credits,
   - the use of regression analysis for analysing the reasons of infringement of the terms of repayment of loans,
   - the use of modelling and forecasting for planning credit activities, and so on.

5. **Economic-statistical analysis of creditworthiness of bank customers.**
   It is the logical continuation of part 4 and it contains the description of the new statistical approaches to the estimation of financial opportunities of the potential customer, that allows to make the grounded decisions about terms of granting credit.

6. **Statistical investigation of liquidity of the balance of the bank.**

7. **Statistical investigation of solvency of the bank.**

8. **Statistical investigation of profitability of the bank.**
   In parts 6-8 the methods of statistical characteristics of three main parameters of bank activities are considered:
   - the description of the tendency of their change;
   - the factors which form these parameters and cause their change;
   - the ground of the difference of the level of indices for regional departments of the bank and their influence on the summary indices in whole in the bank.

   The methods of *Bank Statistics* are based on the data of accounting, managing and financial accounts, in particular – the balance of bank, accounts about the financial results, the forms of inside accounts. But here it is necessary to use also the data of the statistics of the commodity market and market of services, of the statistics of finance, of social statistic, etc.

   It is necessary to stress that in spite of the specific sphere of the use of *Statistics of Bank Activities*, it is a system of knowledge necessary not only for the bank employees, who perform the statistical and analytical work. *Bank Statistics* will be of use, for example, for businessmen in order to estimate in the right way the alternatives with regard to the cost and the quality of the necessary bank services, prices, the terms and the risks when applying for loans. Also, it is known, that when considering the application of the customer for granting him the loan, the bank makes a thorough analysis of his activity. But for businessmen it is very important to realise constantly and independently (but not only in the case of necessity) the inspection of the main indices of the structure of the capital and efficiency with the aim to receive a preliminary self-appraisal of his reliability as the future borrower.
Statistics of Bank Activities as the independent sphere of knowledge has meanwhile not very long history. But has already received very good references from the users as it helps to heighten the efficiency of the activities of banks and the bank system. At the same time such references stimulate the professors, who teach Bank Statistics, to perfect and develop it.

The main trends of development of Bank Statistics both as science and as a course that demand improvement in the nearest future the perfecting are:

- to draw more and more people – theoreticians and practical workers, statisticians and bankers – to use statistical methods in the analysis of bank activities;
- to elaborate in the educational and scientific establishments, analytical centres of banks the themes of scientific researches of bank activities which would require the wide use of statistical methods;
- preparing theses, monographs, articles and reports using the information about the activities of bank system with the grounded use of statistical methods;
- elaborating the curriculum for training bank employees at different levels, using different forms of training (correspondence, distant training, remote-control courses, extern training and so on);
- creating the statistical tools for analysing bank activities of the Central Bank;
- increase of publications about the methods of Bank Statistics with the view to acquaint the potential users with their merits and the necessity to use them while analysing bank activities.
TECHNOLOGICAL LEVEL AND THE CAPACITY TO INNOVATE OF SMALL ENTERPRISES IN POLAND

(Comparison of the Gdańsk and Lublin Provinces)

Andrzej Balicki

ABSTRACT

The present position of small enterprises and their role in the coming years is a combined result of numerous, usually unfavourable factors of both macroeconomic and internal character. The question is whether the importance of the small enterprise sector will increase as it is expected by theoreticians or we will rather observe a great number of failures of companies in this sector.

In this paper we intent to answer the question about the extent to which the small businesses have the capacity to develop and whether their owners are sufficiently entrepreneurial to overcome difficult situations and survive under increasing competition. We assume that the capacity for innovation is an essential feature of entrepreneurship and we focus our attention on this aspect. Additionally we intend to identify differences between particular regions of the country.

Key words: small enterprises, entrepreneurship, innovation,

1. Theoretical introduction

1.1. Innovations and capacity for innovation

Innovation enhancing activities are a series of actions in the area of scientific research, technology and organization, of a financial and commercial character aimed at the development and implementation of new or substantially improved products and processes. Some of those activities are innovative as such;

---

1 This research was undertaken with support from the European Commission’s Phare ACE Programme P97-8123-R.
2 Professor in Department of Statistics, University of Gdańsk, ul. Armii Krajowej 101, 81-824 Sopot, Poland.
some others may not contain the element of newness, but are indispensable for the development and implementation of innovations (Definicje ...., 1999).

The concept of innovation is defined in various ways, however the sense of this concept may always be boiled down to the purposefully designed changes concerning the product (a new or an improved good), methods of manufacturing (new or improved technologies), organization of work and the production process. Innovations may adopt various forms and varieties. Technological innovations (i.e. TPP innovation – technological product and process innovation) may be particularly distinguished among numerous forms of innovations:

- Technological product innovations – these are the innovations whose scope covers both the new and the already implemented products. Product innovation is considered as implemented if a product (or a service) has already been introduced into the market.

- Technological process innovations concern processes, which are a new or are a substantial improvement of the already applied technologies. Implementation means that they are used in the production process.

Both the product and the process have to be new at least from the point of view of the enterprise, which produces or uses it.

The scope of innovation, as it may be derived from its definition, may be much broader. Those activities usually imply organizational changes (so called organizational–technical innovations) which simultaneously stick together the phenomenon of innovation, give it a coherent and logical form. The scope of organizational innovation encompasses a number of elements, starting from the organization of work and production processes, through improvement of logistic systems and the system of staff selection and training to the formation of optimal cost structure. It is difficult to imagine technological innovation without corresponding organizational changes in many areas, supporting it or resulting from it. A particular form of organizational innovation, related with new management techniques is the computerization of the enterprise, the scope of which may be varied, depending on the needs and possibilities.

In view of the definition of innovation enhancing activities, the capacity to innovate should be understood as the capacity and motivation of business entities to continuously search and use in practice new results of scientific research and research and development activities, new concepts, ideas and inventions. The capacity for innovation is aimed to improve and develop the existing technologies concerning production, exploitation, and services, introduction of new solutions in organization and management, improvement and development of infrastructure, in particular concerning the process of acquiring, processing and retrieving

---

1 This is a set of officially approved definitions and methodology of research in the area of innovations, compatible with the international requirements (the so-called Manual Oslo – OECD, Eurostat).
information. The capacity to innovate is thus a feature of business units, while innovation is an effect of certain activities.

1.2. The capacity to innovate, entrepreneurship and development

The above-presented idea of the “capacity to innovate” matches the meaning of the concept of “entrepreneurship” as J. Schumpeter (1960) introduced it. Entrepreneurship constitutes a form of activity that consists in introduction of new combinations of factors of production, introduction of new products and methods, conquest of new sales markets and new sources of supply and creation of new, more efficient forms of business organization. It is then possible to draw only one conclusion from the comparison of those two concepts: entrepreneurship means first of all the capacity to innovate, thus the innovation factors and their effects in the form of innovation are at the same time the factors and manifestations of entrepreneurship. These concepts are, however, not identical. One may have a number of good ideas but avoid the risks, not being able to take advantage of arising opportunities. Entrepreneurship, after P. Drucker (1992) may be understood as a feature (behavior) of the entrepreneur or the enterprise expressed in the readiness and capacity to undertake and creatively solve new problems, the ability to use the appearing chances and opportunities and in flexible adaptation to the changing conditions. Entrepreneurship may be manifested in all areas of activity of the enterprise, e.g. in rapid implementation of technical-organizational progress or in dynamic activities in the market place. “Innovation is the instrument of entrepreneurship” (cf. Drucker, 1992).

Entrepreneurship is often perceived as something connected with small enterprises, their start-ups and management. It seems to be an unjustified limitation of the concept. It is important to say, however, that with the passage of time that elapses from the moment of start-up of the firm, the “entrepreneurial spirit” declines. Innovation exerts influence on the development of enterprise, it is its important driver. The development is then understood as a process taking place in time and resulting in changes of a qualitative nature. It is necessary to distinguish development from growth, which implies quantitative changes only. The development changes the existing status quo, rising it to a higher level. The quantitative growth in output is not necessarily a symptom of development.

The firm’s capacity to innovate is not only a condition of its development, but of its mere survival. All aspects that condition the capacity to innovate determine as well the firm’s development or its possibilities for development. A multisided connection between the capacity to innovate and development may be identified in various aspects of the firm’s operations: the market where it

---

2 This is the most gentle interpretation of the slogan: “innovate or die”.

---
operates (local–national–international), co-operation and co-operative connections, competitiveness, use of computers and the internet, etc.

Small businesses are assigned various functions, that they may perform in the market economy. The fact, that small businesses may be the source of innovations is mentioned as one of the first of those numerous functions (Targalski, 1999). The capacity to innovate of small enterprises may be manifested to a limited extent only, since their research, development and implementation possibilities are usually limited because of their financial capacities. In the most basic sense the capacity to innovate of small enterprises has an organizational, infrastructural character, related to the process of computerization, and to a lesser extent related to new products and technologies.

1.3. The scope of analysis

The analysis of the capacity to innovate as well as the assessment of the innovation processes is a difficult and complex undertaking, since the phenomenon itself has a complex and multisided character. The scale of innovations, intensity of innovation processes, the character of introduced innovations, the length of the cycle of innovation process, barriers inhibiting the introduction of innovations, intentions concerning the innovations in the coming years are the most interesting elements of research.

The research project mentioned above concerns general problems of small enterprises during the process of economic transition in Poland and did not distinguish the issue of innovation in any special sense. However, a number of questions contained in the questionnaire were related to this problem, even though measures usually applied in innovation research were not used. Empirical analysis presented here covers selected aspects of the firm’s development. It is a concept that has not been precisely defined – the definition had a conceptual rather than operational character. It was due to the fact, that development is a complex phenomenon, with a number of various aspects, factors and correlates. It is thus difficult to unequivocally assess the level (or stage) of the firm’s development. It is, however, possible to assess certain characteristic elements of that development. The approach applied in the analysis consisted in the selection of those issues related to the operation of small businesses, which seemed to have been particularly connected with the concept of development. Particular attention has been devoted to the technological level of enterprises and their capacity to innovate. The questions concerning those issues might have been found in

1 The research of GUS (Korona, 1994) revealed the rule, that the larger the enterprise, the higher its activity and intensity of innovations. This results i.a. from higher financial possibilities of large companies, easier access to information or the possession of R&D facilities.

2 The most advanced research of innovations is carried out by the Main Statistical Office (GUS) (e.g. questionnaire survey of the enterprises’ capacity to innovate in 1992 on a sample of 2.5 thousand units – small, medium-sized and large enterprises).
different sections of the questionnaire applied in the project. They form a combined description of the capacity to innovate.

It is however always interesting to know what is the current technological level of small enterprises. It is also the launching pad for a number of activities, but at the same time the source of potential and real troubles and difficulties.

2. The empirical section

2.1. The technological level of small enterprises

Certain questions of the survey concern the technological level of the firm in sphere of its processes and products (goods or services) as well. Both spheres perform two roles here: the symptoms of the capacity to innovate and at the same time the background for its assessment. One should not expect the enterprise to achieve a high technological level without carrying out innovation-oriented activities. In the population of enterprises operating in different sectors of economy it is difficult to find a uniform way of measuring the absolute technological level. Thus the authors relied on self-assessment, assuming that the firms (their owners, managers) have a good orientation in the area of contemporary technologies of processes, products and services so that they can without much ado indicate the position of their enterprise on a simple scale of “technological advancement”. It has necessary been a relative assessment. We cannot derive from the answers, what and when has been implemented or changed – neither as far as the product nor as the process is concerned. The responses are the firm’s own subjective assessment of the enterprise, what are its current capacities or what is the level of technology it applies. The scale used to express the opinions of those levels of technological advancement has 5 degrees (very low, low, average, high, very high). For operational purposes the values of 0, 1, 2, 3, 4, 5 were correspondingly assigned to particular degrees of the scale.

Both partial 5-degree scales were combined together. Their cross-combination gives 25 classification classes. The contingency table resulted from the combination of both scales reveals first of all the correlation of assessments in both ranges that are of interest to us (cf. Table 1 and 2). However, in the last row and in the last column of such a table there are only one-dimensional (marginal) distributions of “advancement” in relation to each scale separately.

Table 1. Technological level of small firms in the Province of Gdańsk

<table>
<thead>
<tr>
<th>Technological level of the process ((X_1))</th>
<th>Technological level of the products ((X_2))</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>very low (0)</td>
<td>low (1)</td>
<td>average (2)</td>
</tr>
<tr>
<td>low (1)</td>
<td>high (3)</td>
<td>very high (4)</td>
</tr>
</tbody>
</table>

1 The survey, which has yielded data for analyzing small enterprises’ problems in transition period in Poland was characterized by Szreder (2001).
### Table 2. Technological level of small firms in the Province of Lublin

<table>
<thead>
<tr>
<th>Technological level of the process ($X_1$)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>very low (0)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>low (1)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>average (2)</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>10</td>
<td>1</td>
<td>59</td>
</tr>
<tr>
<td>high (3)</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>71</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>very high (4)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0</td>
<td>1</td>
<td>50</td>
<td>81</td>
<td>2</td>
<td>134</td>
</tr>
</tbody>
</table>

$\chi^2 = 2289.98; \quad p = 0.00000 \quad \overline{x}_1 = 2.54 \quad \overline{x}_2 = 2.63$

See note under Table 1

In both provinces the assessment of technological level is quite well balanced and oscillates between an average and a high level (2.54 – 2.86). We can assess this level as above average, but still not high. The companies assess the technological level of their products of services somewhat higher than the technological level of their processes. Both assessments are somewhat higher in the Gdański than in the Province of Lublin. The survey reveals, that small enterprises in Poland do not have – on average – modern production facilities or the most modern technologies at their disposal, but it does not support the widespread opinion about their very low technical and technological level either. From a survey of the Central Statistical Office in Warsaw (GUS) one can infer, that in the years 1990-1992 at least 50% of small enterprises in the industry introduced some innovations, at least 38% of them introduced product innovations and at least 20% introduced innovation in their processes. It had without doubt a positive influence on the present general level of technological
advancement. It is, however, not a level, which is sufficiently high to eliminate
the concern of the existence of a production barrier or to forget fears about the
competitiveness of the small business sector in Poland (cf. Piasecki, 1998). In
view of the integration with the European Union this is a real threat for existence
of a large part of small enterprises.

In both provinces the technological levels of processes and products are
mutually dependent. It manifests itself by the numbers accumulated along the
diagonal of the contingency table, and very high values of the chi-square statistics
(we definitely reject the hypotheses of independence of those levels). Thus we can
concluded that the technological level of products is not formed independently
from the technological advancement of the process. Higher assessment of one
value corresponds to the average higher assessments of the other value.

2.2. Joint self-assessment of the technological level

Independently from the partial description of each range of the technological
level, one combined scale of "technological advancement" has been created on
the basis of two 5-degree (from 0 to 4) partial scales: the self assessment of the
technological level of process and the technological level of product.

Assuming, that the scales are not independent and are positively correlated, a
combined scale of technological advancement (TA) was construed according to
the following procedure. The correlation coefficient \( r_{jk} \) between partial scales
was computed, assuming at the same time that \( r_{jk} = \cos \alpha_{jk} \), where \( \alpha_{jk} \) is the
angle of vectors representing particular \( j \) and \( k \) scales. That angle was
consequently determined from the correlation coefficient \( \alpha_{jk} = \cos^{-1}(r_{jk}) \), and
then it was possible to determine \( \tan(\alpha/2) \). If \( |r| = \cos \alpha > 0 \), then the values on
the transformed (combined) scale were calculated from the following formula:

\[
y = 2x \tan(\alpha/2)
\]

in case when the values on the scale were in agreement, i.e. for \( x_1 = x_2 = x \), e.g.
(1,1), (2,2), (3,3), (4,4), or

\[
y = (x_1 + x_2) \tan(\alpha/2)
\]

in the opposite case (\( x_j \) is the result on the partial (variable) scale \( j \)).

A particular case of that procedure is \( y = x_1 + x_2 \) for non-correlated scales
(\( r = \cos \alpha = 0 \); \( \alpha = 90^\circ \); \( \tan(\alpha/2) = 1 \)), what suggests an additive character of the
scales.

In order to normalize the value of the new scale (so that it takes values from
0 to 1) a quotient transformation was applied

\[
y^* = y/k
\]
where \( k \) is the maximum value of the new scale, assuming that the initial scales are not correlated (in our case its value was equal to 8).

Creation of a combined scale of the technological level of small enterprises requires estimation the angle of vectors from the coefficient of correlation between the following variables: “technological level of the process” and “technological level of the products”. For the Province of Gdańsk we get:

\[
r(X_1, X_2) = 0.6749, \text{where from } \alpha = 47.55^0 \text{ and } \text{tg}(\alpha/2) = 0.4406
\]

and for the Province of Lublin:

\[
r(X_1, X_2) = 0.7961, \text{where from } \alpha = 37.24^0 \text{ and } \text{tg}(\alpha/2) = 0.3369
\]

It is then possible to create the coefficient of technological advancement (a new variable \( X_3 \)) according to the following formula

\[
TA = \frac{(X_1 + X_2) \text{tg}(\alpha/2)}{8}
\]

One can easily see that the values of the scale for all surveyed units are proportionally changed in a relation equal to \( \text{tg}(\alpha/2)/8 \), while higher correlation results in a more substantial correction of the sum of the \( X_1 \) and \( X_2 \) variables. Higher correlation of the technological level of the product and the process means that one magnitude determines a larger part of variability of the other magnitude, then other factors exert a smaller influence on them, particularly on the technological level of products and services provided (e.g. the human factor, good organization of production processes, strict compliance with the technological regime, quality control). Thus the “value” of particular elements of technological level will be higher in the case of lower correlation. This line of reasoning should not arise more doubts in case of positive correlation of both magnitudes.

The distributions of indices of technological advancement in the studied provinces are given in Table 3. Because of the possible values of the \( X_1 \) and \( X_2 \) variables (0,1,2,3,4), the \( TA \) variable may adopt only 9 different values, including the value of 0.

The distributions of \( TA \)-variable in both provinces have the same shape (reflecting their bimodal character), but different location. In our opinion this changed location more appropriately represents their real technological level than a simple, additive scale. We have thus a situation where even if in each separate dimension the technological level was assessed as above average, then the combined technological level would be assessed as relatively low. There is practically no small enterprise that is seriously technologically advanced.

**Table 3.** The distributions of technological advancement of small enterprises

<table>
<thead>
<tr>
<th></th>
<th>The Province of Gdańsk</th>
<th>The Province of Lublin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Province of Lublin looks in a different way in comparison with the Province of Gdańsk (cf. Table 3), even though it is difficult to say that its situation is much worse.

Let us return then to the values of tangens, since these are the factors that correct the sums of values on the $X_1$ and $X_2$ scales. One could expect the correlation of values of the $X_1$ and $X_2$ variables. Given the number of observations (213 and 134) it is quite high. However, a higher correlation reduces the importance of the values of both combined variables into a new scale. It is then difficult to unequivocally – more negatively - assess the firms from the Province of Lublin, for which the corresponding sum of two measures of $X_1$ and $X_2$ is more strongly corrected downwards, in average by as much as 23.5% (the quotient of $0.4406/0.3369$). This is due to the fact, that stronger correlation of both combined variables results in higher share of variability of one variable explained by the second, thus both variables convey more the same information in the Province of Lublin ($r^2 = 0.634$) than it is in the case of the Province of Gdańsk ($r^2 = 0.455$).

### 2.3. The capacity of small enterprises to innovate

The study of innovation activity and its effect in the form of innovations is the principal aim of the present paper. A direct question was put in the survey: whether in the years 1998-1999 there were innovations taking place in the firm in relation to the product (new or improved goods), in the area of process, or in the area of organization (e.g. in the way of managing the enterprise). The capacity to
innovate is a multisided concept. It is determined by various activities, behaviours, facts etc. Three aspects of the capacity to innovate were distinguished in this question. Each firm is thus characterized with three constituting elements of the capacity to innovate. In each of those aspects the firm may be assigned the value of 1 if a given element appeared in the years 1998-99 or the value of 0 if it did not appear. We receive then a combined table (cf. Table 4) which presents the scope (type) and range (no. of firms) of the capacity to innovate.

Table 4. The capacity to innovate of small firms in the Gdańsk and Lublin Provinces

<table>
<thead>
<tr>
<th>Sphere of innovations</th>
<th>Province of Gdańsk</th>
<th>Province of Lublin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes (1)</td>
<td>No (0)</td>
</tr>
<tr>
<td>in the product</td>
<td>102 (49.3%)</td>
<td>105 (52.2%)</td>
</tr>
<tr>
<td></td>
<td>41 (29.9%)</td>
<td>96 (70.1%)</td>
</tr>
<tr>
<td>in the process</td>
<td>30 (14.6%)</td>
<td>176 (85.4%)</td>
</tr>
<tr>
<td></td>
<td>22 (16.1%)</td>
<td>115 (83.9%)</td>
</tr>
<tr>
<td>in the organization</td>
<td>43 (20.8%)</td>
<td>164 (79.2%)</td>
</tr>
<tr>
<td></td>
<td>26 (19.0%)</td>
<td>111 (81.0%)</td>
</tr>
<tr>
<td>In all three spheres *</td>
<td>3</td>
<td>55 (100%)</td>
</tr>
</tbody>
</table>

*"In all three spheres” – in the last row of the Table– these are the numbers of firms which were assigned the value of 1 or 0 in all three spheres of their activity simultaneously.

We can immediately notice in Table 4 the little percentage of small enterprises which try to implement the technological innovation (in the sphere of process innovation) constituting merely ca. 15% of the total firms in question, accompanied by a considerable percentage of innovative enterprises in the sphere of product innovation. It is then possible to consider, that small enterprises quite readily undertake innovative activities in the product sphere, most probably because they are less capital intensive, and they perceive their chance to cope with competition or clients’ requirements in this area. The innovations in the sphere of technical production equipment and provision of services constitute quite a substantial problem for small businesses. And even if we do not analyze the structure of innovations further, it seems that they “do not go hand in hand”; it is impossible, however, to exclude that the technological innovations preceded the product innovation and had been implemented prior to the year 1998. This is, however, rather unlikely.

The analysis of data indicates as well, that the percentage of small enterprises that could be described as innovative is higher in the Province of Gdańsk than in the Province of Lublin. It is then necessary to study the relationship between the technological advancement and innovation activities of firms undertaken in the years 1998-99 (see comment to the subsequent items). To this end it is possible to create one combined scale of the capacity to innovate or to study a number of various correlations.
To make the analysis more comprehensive it is possible to determine the strength and the direction of associations between different kinds of innovations, i.e. the relationship between the presence or absence of its particular elements (cf. Kendall and Buckland, 1975, Yule and Kendall, 1966). The three distinguished elements and their presence or absence may be distributed among three tables of associations, studying the dependence of presence of particular constituting pairs of innovations. Omitting details concerning the associations, we would like to inform here that in each of three distinguished cases (process – product, process – organization, product – organization) we notice the so-called negative association only. It means that innovations in two spheres occur together less often (simultaneous declaration: yes – yes) and at the same time lack of innovations occur more often (simultaneous declaration: no – no) than expected if independent. The simultaneous negative declaration means, however, something entirely different than simultaneous positive declaration and indicates a supposed absence of the capacity to innovate. We would thus consider this result as negative from the merit point of view, even under condition of formal identification of dependence. Thus, in the Province of Gdańsk the percentage of non-innovative small firms as far as process and product is concerned amounts to 40.3%, while in the Province of Lublin to as much as 54.7% (cf. Table 5).

Table 5. Percentage of non-innovative firms in the Gdańsk and Lublin Provinces*

<table>
<thead>
<tr>
<th>Lack of innovations in the sphere of</th>
<th>the product</th>
<th>the process</th>
<th>the organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>the product</td>
<td>×</td>
<td>40.3%</td>
<td>36.2%</td>
</tr>
<tr>
<td>the process</td>
<td>54.7%</td>
<td>×</td>
<td>67.0%</td>
</tr>
<tr>
<td>the organization</td>
<td>51.1%</td>
<td>66.4%</td>
<td>×</td>
</tr>
</tbody>
</table>

Above the diagonal – the Province of Gdańsk, below the diagonal – the Province of Lublin

One could expect such proportions of simultaneous existence of various kinds of innovation on the basis of analysis of the arrangement of numbers in Table 4 (small numbers of small enterprises manifesting particular types of innovation favour the appearance of correspondingly smaller numbers of enterprises manifesting innovation in two spheres at the same time). For instance in the Lublin Province there are 41 (29.9%) small enterprises, which declare technological innovation in the product sphere and 21 (16.1%) enterprises, which declare technological innovation in the process sphere, but there is only 1

1 Association is the degree of relationship between two dichotomic features in a 2×2 table.
enterprise, which manifests innovations in both spheres. The same, even though non-identical situation prevails in two provinces taken for comparison: i.e. the Gdańsk and the Lublin. Negative association of particular types of innovations is quite astonishing and indicates that in small enterprises innovations tend to be one-dimensional and, what is even more striking – mutually exclusive. The only positive explanation of this situation is the distribution of innovative activities in a longer time period (if they are implemented at all).

The way to quantify the associations between particular types of innovation is to calculate the point correlation coefficients. Using the data about the existence and co-existence of innovation we calculate them according to the following formula

$$\phi_{jk} = \frac{n \cdot n_{jk} - n_{jj} \cdot n_{kk}}{\sqrt{n_{jj} \cdot n_{kk} (n-n_{jj})(n-n_{kk})}}$$

where $n$ denotes the number of surveyed enterprises, $n_{jk}$ – the number of enterprises, where two forms of innovations - $j$ and $k$ - co-exist, while $n_{jj}$ and $n_{kk}$ denote the number of enterprises, where innovation of the $j$ type coexist with the innovations of the $k$ type, correspondingly.

Thus we obtain a matrix of point correlation coefficients for the Province of Gdańsk

$$\Phi_{Gda} = \begin{bmatrix} 1 & -0.1611 & -0.1981 \\ -0.1611 & 1 & -0.0766 \\ -0.1981 & -0.0766 & 1 \end{bmatrix}$$

and, correspondingly, for the Province of Lublin

$$\Phi_{Lub} = \begin{bmatrix} 1 & -0.2424 & -0.3163 \\ -0.2424 & 1 & -0.1103 \\ -0.3163 & -0.1103 & 1 \end{bmatrix}$$

which corroborate the previous results suggesting weak or negative correlation.

Not each correlation coefficient that is different from 0 has to suggest the existence of statistically meaningful correlation. The chi-square independence test eliminates any doubts in this area. The results of testing are presented in a combined table (6 and 7), where we give the observed values of the chi-square statistics in the sample over the diagonal, while under the diagonal the values of probabilities corresponding to them (the so-called $p$-values) are given.

Table 6. Relationships between innovations in small enterprises in the Province of Gdańsk

<table>
<thead>
<tr>
<th>Innovations in the sphere of</th>
<th>the product</th>
<th>the process</th>
<th>the organization</th>
</tr>
</thead>
</table>

Innovations in the sphere of the product: $\times$; $\chi^2 = 5.3493$; $\chi^2 = 7.8742$

<table>
<thead>
<tr>
<th>Innovations in the sphere of</th>
<th>the product</th>
<th>the process</th>
<th>the organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>the product</td>
<td>$\times$</td>
<td>$\chi^2 = 8.0513$; $\chi^2 = 13.7051$</td>
<td></td>
</tr>
<tr>
<td>the process</td>
<td>$p = 0.00455$</td>
<td>$\times$</td>
<td>$\chi^2 = 1.6662$</td>
</tr>
<tr>
<td>the organization</td>
<td>$p = 0.00021 \times$</td>
<td>$p = 0.1968$</td>
<td>$\times$</td>
</tr>
</tbody>
</table>

* Significance does not raise any doubts.

Table 7. Relationships between innovations in small enterprises in the Province of Lublin

Innovations in the organizational sphere deserve some comment. It seems that the organizational changes as innovations have a somewhat different caliber than technological changes or improvement in the quality of goods and services. It seems that they do not need to appear at the same time as the technological changes, unless the latter ones have a revolutionary character. Let us firstly consider product innovations and organizational innovations. The assessment of parallel appearance of both spheres of innovation is similar to the former case – there is a strong negative association. Total negative association of both phenomena in the Province of Lublin is particularly apparent. It means that there are no small enterprises where one could simultaneously observe the capacity to innovate in the product sphere and in the organizational sphere, even though they exist separately, as if independent one from another among many of them. Summing up, there exists no statistically significant relationship in the product sphere and in the organizational sphere, both among the enterprises of the Province of Gdański and of the Province of Lublin. A typical negative association “does not result in positive” assessment of the capacity to innovate.

2.4. The index of the capacity to innovate

Since the coefficient of point correlation is the Pearson correlation coefficient for alternative variables, too low or negative values of those coefficients exclude the possibility to construct a combined and reasonable measure of the capacity to innovate in the form of the 1st principal component. In case of low correlation coefficients the 1st principal component does not have a dominant importance (low intrinsic value and low share in explanation of the total variability) and cannot be considered as an overall index.
We propose then to apply another approach. We have distinguished three forms of innovations. These are relatively independent (at least in the case of small enterprises) magnitudes characterizing three types of innovations. We will combine them into one **innovation scale**. The best way to do it in the simplest form is to sum up the values (equal to 0 or 1) assigned to particular objects on the basis of three different features – forms of innovations \((j = 1, 2, 3)\) considering them as equally important. In our opinion the attempt to assign to them different weights could change the image of innovations in an unjustified way (since we do not know the details of those innovations) even though it could surely make the scale more attractive. Thus we will be able to assign the value equal to the sum of numbers equal to 1 to a particular enterprise. We will thus receive a four-point scale of the capacity to innovate with the possible values of 0,1,2,3. It is possible to normalize that scale quite simply dividing particular results by the highest possible result, i.e. 3. Such “index of the capacity to innovate” will thus have the following form:

\[
CI_{(i)} = \frac{\sum_{j=1}^{p} x_{ij}}{p}
\]

where \(x_{ij}\) is the value of the \(j\)th feature in the \(i\)th enterprise (1 or 0), while \(p\) is the highest possible sum of results, in this case being equal to the number of features.\(^1\)

We will consider that index as the **index of the capacity to innovate** \((CI)\).

Table 8 presents the distribution of the index of the capacity to innovate of small enterprises in both surveyed provinces. One should bear in mind that in our study this problem concerns only two years, i.e. 1998 and 1999.

The interpretation of the index is quite simple. The value of 0 – no innovations in all spheres, it means that the enterprise is not innovation oriented; the value of 0.33 – the capacity to innovate was noted in one sphere, it means that the enterprise is characterized with low and selective capacity to innovate; the value of 0.66 – there were innovations in two out of three spheres, meaning that the enterprise manifests a moderate capacity to innovate and the value of 1.00 – innovations are manifested in all three spheres, or the enterprise manifests a high capacity to innovate.

**Table 8. Index of innovations in small firms**

<table>
<thead>
<tr>
<th>Value of the index</th>
<th>Province of Gdańsk</th>
<th>Province of Lublin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of firms</td>
<td>Percentage share of firms</td>
</tr>
<tr>
<td>0.00</td>
<td>55</td>
<td>26.70</td>
</tr>
</tbody>
</table>

\(^1\) The scale may be made more “complex” introducing certain weights taking into consideration mutual correlations between the variables combined.
The picture of the capacity to innovate that emerges from the analysis of the index suggests that the Province of Gdańsk is slightly more advanced in this respect and confirms the results presented in Table 5. A question, however, appears how this result relates to the theoretically high flexibility of small and medium-sized enterprises in reacting to the needs of the market, enhancing the capacity to innovate of that sector of economy? (compare Piasecki, 1998, p.177-178).

2.5. Innovations and the level of technological advancement

Total association (per analogy to the correlation) concerns two dichotomic features only. No attention is paid to the existence of any information about other features of the population of objects. If such information is available, then it is possible to isolate particular class of objects having some third feature (and some further features) and study the association of the principal features. Such approach to research has a considerable weight in the statistical practice and allows to identify factors (causes, intermediary variables) of total associations concerned, which are sometimes of an illusionary character. This approach is particularly justified if the surveyed populations are not uniform, constitute a mixture of objects, of which some have, and some do not have an intermediating feature. The research then boils down to the study of associations in sub-populations. Such associations are known as partial associations (see: Yule and Kendall, 1966).

Using that approach the population of small enterprises was divided into two categories: those manifesting low level of technological advancement (lower than the average level – cf. Table 3) and manifesting a high level of technological advancement (higher than the average level). One has to bear in mind, that in comparable provinces there are different average levels of technological advancement. The associations of particular types of innovations were studied within those groups. Thus a new feature was included into the analysis and actually four features are being analysed: the capacity to innovate in three spheres and the level of technological advancement. We confine our analysis to certain associations only. The following results were obtained (analytical tables are omitted here):

1. Among the small enterprises characterized by a “low” level of technological advancement in the Province of Gdańsk one can observe a nearly “textbook” independence of particular types of innovations (as far as product, process and organization is concerned). It means that even if the innovation-oriented
activities are observed, they have an incidental, chaotic character. On the other hand among the small and technologically weak enterprises of the Province of Lublin one can observe some signs of dependency (between the product, process and organizational innovations). This association is, however, completely negative, i.e. there are no enterprises which manifest innovations from the two correlated spheres at the same time.

2. Among the small enterprises of “high” level of technological advancement one can observe a statistically significant negative association between particular types of innovations: between the product and process innovations \((\chi^2 = 9.271; p = 0.002328)\); and between the product and organizational innovations \((\chi^2 = 8.180; p = 0.004236)\). The association between the process and the organizational innovation is close to the level of significance \((\chi^2 = 2.865; p = 0.090524)\). The fact, that the associations are negative, suggests the inconsistency of innovation-oriented activities (undertaking of certain activities) and about much more frequent cases of the failure to undertake any activities than undertaking them. A similar situation seems to prevail among the small and “technologically advanced” enterprises of the Province of Lublin. The statistically significant relationship exists only between the product and process innovations \((\chi^2 = 7.624; p = 0.00575)\); between the product and organizational innovations \((\chi^2 = 7.624; p = 0.00576)\), and is “nearly significant” in the third sphere, that is between the process and the organizational innovations \((\chi^2 = 3.676; p = 0.05521)\). It does not change, however, the generally negative assessment of the innovation oriented activities in this group of enterprises, since the associations are totally or highly negative. In this sense both Provinces are similar to one another, although in the Province of Gdański we find enterprises which undertake innovation activities at least in two spheres at the same time more frequently than in the Province of Lublin (but not too often).

Quite an interesting and complementary picture may result from the question about the relationship between the innovation and the level of technological advancement. Using the analysis of associations an astonishing lack of relevant association between the level of technological advancement (low – high), and any type of innovation among the small enterprises of the Province of Gdański has been observed. It means, that the percentage of innovative enterprises in particular spheres is nearly the same among the enterprises that are highly technologically advanced and among the enterprises characterized with a low level of technological advancement. That problem looks somewhat differently in the Province of Lublin. There is there a higher share of firms manifesting innovation-oriented activities among the enterprises with a high level of technological
advancement than among those, which are less technologically advanced. One relationship is statistically significant (between the technological level and the organizational innovation: $\chi^2 = 4.4999; p = 0.0339$), one is close to the statistical significance (between the technological level and the capacity to innovate in the sphere of process innovation: $\chi^2 = 2.885; p = 0.08939$), while the third is insignificant (between the technological level and the product innovation orientation). This comparison does not suggest at all, that the percentage of innovative enterprises (in each sphere) or the percentage of more technologically advanced enterprises is higher in the Province of Lublin, because this is not the case. It only means that in spite of absolutely worse situation of small enterprises in that province, some proportions tend to adopt a better shape, what may result in an improvement of the situation and reduce the gap dividing it from the Province of Gdańsk. Generally speaking, the revealed relationships do not suggest that small enterprises in both provinces have a favourable future, unless their strategy of development is changed.

The lack of relationship or weak relationships between the innovations and the level of technological advancement is confirmed with the correlation coefficients between the proposed scales. The correlation coefficients between level of technological advancement ($X_3$) and the capacity to innovate ($X_7$) are equal to $r = 0.154$ for the Province of Gdańsk, and to $r = 0.1013$ for the Province of Lublin.\footnote{The correlation coefficients were calculated for full pairs of values of both variables (pairwise deletion).}

**REFERENCES**

Definicje pojęć z zakresu statystyki nauki i techniki (1999), Zeszyty Metodyczne i Klasyfikacje, Główny Urzad Statystyczny, Warszawa.


\text{After: Bartnicki M. i Kulikowska M., Sukcesy i porażki w królestwie}


TARGALSKI, J (red.) (1999), Przedsiebiorczość i rozwój firmy, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.


SOME DATA QUALITY ISSUES
IN STATISTICAL PUBLICATIONS IN POLAND

Jan Kordos

ABSTRACT

This paper focuses on the reporting and presentation of information on sources of error in several dissemination media: short-format publications, main report from the surveys, analytic publications, and the Internet.

This review is based on the work of the author conducted for training purposes and in order to help characterise current practices for reporting sources of error in publications. The first study reviewed press releases and short publications of several pages (“short-format reports) issued by six divisions of the Central Statistical Office (CSO): (i) labour, (ii) demography (iii) living conditions, (iv) agriculture, (v) enterprise, and (vi) trade & services. The second study reviewed main reports from the sample surveys published by the same divisions. These are printed reports of methodology of the survey, descriptive analysis, and tables with main results from the survey. The third study reviewed selected “analytic publications” from the same divisions. The fourth review is connected with the Internet which has become one of the media of dissemination of statistical data from the CSO since 1997. Handbooks and articles of methodological nature devoted to some aspects of data quality, e.g. presentation and interpretation of sampling errors, coverage errors, non-response errors, measurement errors are also being considered.

Key words: data quality, sample survey, sampling error, nonsampling error, coverage error, measurement error, processing error, analytic publication, error sources.

1. Introduction

Since 1989 the Central Statistical Office of Poland (CSO) has started adjusting official statistics to international standards, and made the attempts to harmonise it with the European Union statistics. Basic nomenclature and classifications have been adopted, the programme of surveys was considerably changed, new surveys were launched, new methods were applied and the

1 Central Statistical Office/ University of Ecology and Management, Warsaw, Poland.
methodology adjusted to the requirements of market economy and European standards.

Most of the complete statistical reporting stopped and new sample surveys started, such as labour force survey, agricultural surveys, business surveys, enterprise surveys, and several ad hoc sample surveys connected with living conditions, health care, time use, etc. Household budget survey has been adjusted to new requirements. However, the CSO has not yet developed quality guidelines, like Statistics Sweden (Lyberg, Biemer and Japec, 1998) or Statistics Canada (1998), and quality issues of these surveys are important problems to be solved. Therefore, there are different approaches to quality treatment in the different divisions of the CSO. It is possible to observe these different approaches to quality treatment in the surveys conducted by the CSO in information on data quality included in the publications for users.

It is generally accepted that users of survey data need information about a survey’s quality to assess its results properly. It is a very well known fact that there are many dimensions to survey quality and the measurement and presentation of this information is no easy task. Report formats, dissemination media, agency policy and practices vary. The Polish CSO has not much experience in this field.

The previous statistical system was based mainly on complete reporting, simple questionnaires and formal instructions, simplified tables and limited descriptive analysis. In the majority of cases, it did not require modern methods of data collection, designing of sophisticated questionnaires and advanced training of the staff conducting surveys, sophisticated methods of control of in-field operations. Thus it has now become indispensable to organise an extensive training of statistical staff to upgrade their qualifications to make them able to conduct modern statistical surveys. These new skills are acquired gradually.

The pressure from users on quality as expressed in users’ satisfaction surveys, focuses mainly on timelines and availability, which does not encourage specific investigation in accuracy. Among public statisticians there is also a strong feeling that it is difficult to make quality understandable to non-specialists, and it is the professional responsibility of statisticians to decide on necessary quality levels (Kasprzyk et al, 1999).

To help prevent misunderstanding and misuse of data, full information should be available to users about sources, definitions, and methods used in collecting and compiling statistics, as well as their limitations.

The policy of openness in providing full descriptions of data, methods, assumptions, and sources of error is one that is accepted by the CSO (Walczak, 1999) and is connected with the Polish Statistical Law. However, as it can be seen from the practice of different divisions, the implementation of the policy can vary in many ways.

---

The sources of error that affect survey data quality – sampling error, coverage error, nonresponse error, measurement error and their measurement are described in a number of handbooks and articles published in Poland (e.g. Bracha, 1996a, 1996b, 1998; Kordos, 1985, 1988, 1994, 1995, 1996a, 1996b, 1998; Zasępa, 1969, 1972, 1979, 1993). This paper focuses on sources of error in several dissemination media (short-format publications, main reports from the surveys, analytic publications, and Internet). This review is based on the work of the author prepared for the training programme of statistical staff. Before presenting some aspects of data quality in different kinds of publications with survey results, a short presentation of sample surveys conducted by the CSO in last ten years is given.

2. Sample surveys conducted by CSO of Poland in 1991-2000

The experience gained in sampling surveys under the previous system was useful in transition period only in a limited scope. New problems and tasks emerged which were not known to our statistics before. Among them the following should be mentioned here (Kordos, 1996a):

1. Building confidence in statistical information system in the society, as a precise and useful instrument for the description of the social and economic environment.
3. Satisfying the data needs of decision-makers both in the private and public sectors.
4. Planning of statistical programmes in such a way as to reduce the response burden and facilitate establishing systems which make use of sample surveys and reduce bureaucratic procedures.

In the period of transformation new problems emerged which were unknown before. These are mainly the following:

a) use of administrative registers for statistical purposes,
b) integration of data from various sources,
c) development of corresponding systems for national accounts,
d) development and implementation of registers of employers,
e) training of experts in the use of various international standards and classifications and adaptation of those classifications to the national conditions,
f) extension of sample surveys to a broader scale, especially in the economic statistics (designing of questionnaires, training of interviewers, etc.),
g) extension of the methods compensating non-response (methods of weighting the results, imputations, model approach, simulation, etc.).
In this short report it is impossible to present methodology of sampling surveys in detail. General description starting with the name of the survey, its type, sample size and response rate will only be presented.

Before 1990, sampling methods were mainly used in household budget surveys (Kordos, 1985, Lednicki, 1982), agricultural surveys (Kordos, Kursa, 1997), demographic surveys, and mainly to speed up tabulation of collected information in population censuses (in 1950, 1960 and 1970) and in sample censuses (microcensuses) of population (in 1974 and 1984) (Zasepa, 1993).

After 1990 the household budget survey (HBS) was adjusted to new requirements (Central Statistical Office, 1999; Kordos, 1996b) and in 1992 for the first time a new survey of labour force, i.e. labour force survey (LFS) was introduced (Szarkowski and Witkowski, 1994). Other new sample surveys of households were launched such as survey on the well-being of households, health status of households and multi-aspect survey of the living conditions of the population (Kordos, 1998).


With LFS subsamples the following modular surveys were carried out: socio-economic status of the unemployed (August 1993), rural labour market (November 1993), effectiveness of labour market policy (August 1994 and 1996), professional career of the graduates (November 1994 and 1997), the situation at the labour market and the living conditions of the disabled (February 1995), unregistered labour (August 1995 and 1998), disabled persons on labour market (1995, 2000). It should be stressed that since the fourth quarter of 1999 the LFS has been carried out as a continuous survey. The quarterly sample currently amounts to 24,440 dwellings. It was constructed in such a way that every one of 13 weekly samples is not only the same size but has also the same structure. Selection of quarterly samples is performed according to the rotation system: 2 (2) 2, i.e. two quarters in the sample, two quarters out of the sample and two quarters again in the sample. The results are processed and published quarterly. Results of the survey conducted through this method allow presentation of situation on the labour market during the whole quarter (Central Statistical Office, 2000).

Table 1. The sample surveys conducted by the Central Statistical Office of Poland in 1991-2000
<table>
<thead>
<tr>
<th>Serial Number</th>
<th>Sample Survey</th>
<th>Year(s)</th>
<th>Type of the survey</th>
<th>Sample size (in thous.)</th>
<th>Response rate (in percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.</td>
<td>Survey of the level of wages and salaries by occupation</td>
<td>1998,1999</td>
<td>Periodic</td>
<td>23.7</td>
<td>80-78</td>
</tr>
<tr>
<td>5.</td>
<td>Survey of the employed by the level of wages and salaries</td>
<td>1991-1997</td>
<td>Periodic</td>
<td>18.0-64.0</td>
<td>80</td>
</tr>
<tr>
<td>7.</td>
<td>Small-Sized Enterprise Survey</td>
<td>1995-2000</td>
<td>Periodic</td>
<td>100.0</td>
<td>40</td>
</tr>
<tr>
<td>8.</td>
<td>Land Use, Sown Area and Livestock Survey</td>
<td>1991-2000</td>
<td>Periodic</td>
<td>120-90</td>
<td>95</td>
</tr>
<tr>
<td>11.</td>
<td>Interview Crop Production Survey</td>
<td>1991-2000</td>
<td>Periodic</td>
<td>30.0</td>
<td>98</td>
</tr>
<tr>
<td>13.</td>
<td>Multi-aspect survey of the living conditions</td>
<td>1997</td>
<td>One-time</td>
<td>12.5</td>
<td>86</td>
</tr>
<tr>
<td>15.</td>
<td>Young Couple Survey (second round)</td>
<td>1995</td>
<td>Panel</td>
<td>9.6</td>
<td>58</td>
</tr>
<tr>
<td>16.</td>
<td>Monitoring of Living Conditions</td>
<td>1995,1996</td>
<td>Panel</td>
<td>2.7-2.3</td>
<td>87</td>
</tr>
<tr>
<td>17.</td>
<td>Sample Census of Agriculture</td>
<td>1994</td>
<td>One-time</td>
<td>206.0</td>
<td>n.a.</td>
</tr>
<tr>
<td>19.</td>
<td>Consumer Tendency Survey</td>
<td>1997-2000</td>
<td>Cont./Q</td>
<td>2.3</td>
<td>86-76</td>
</tr>
<tr>
<td>20.</td>
<td>Innovation Survey in Services</td>
<td>1997-2000</td>
<td>Periodic</td>
<td>17.5</td>
<td>70</td>
</tr>
<tr>
<td>21.</td>
<td>Retail Shops Activity Survey</td>
<td>1997-2000</td>
<td>Cont.</td>
<td>18.7</td>
<td>77</td>
</tr>
<tr>
<td>22.</td>
<td>Health Status Survey</td>
<td>1996</td>
<td>One-time</td>
<td>22.2</td>
<td>88</td>
</tr>
<tr>
<td>24.</td>
<td>Post-Enumeration Survey - Census of Agric.</td>
<td>1996</td>
<td>One-time</td>
<td>11.4</td>
<td>96</td>
</tr>
<tr>
<td>25.</td>
<td>Time Use Survey</td>
<td>1996</td>
<td>One-time</td>
<td>1.0</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

n.a. - not available

New sample surveys of the employed in the national economy started. At this point we will mention only two sample surveys of the employed:

(a) the survey of the employed by the level of wages and salaries, and
(b) the survey of the level of wages and salaries by occupation.
New sample surveys started in economic statistics: labour costs survey, business tendency survey in industry, construction and trade (Barczyk, 1995), small enterprise survey, innovative survey (Niedbalska, 2000), and other enterprise surveys (qualitative and quantitative nature) (Zagozdzinska, 1996; Barcikowski et al., 1999). More important sample surveys conducted in 1991-2000 are presented in table 1. There are names of the sample surveys, year or years when surveys were carried out, their types (continuous, periodic or one-time), size of the sample, and approximate response rates in percent.

3. Reporting Sources of Errors in Dissemination Media

There are many formats available for release of results from CSO data collection programmes, ranging from press release and short format reports of several pages to detailed and sophisticated analytic reports. Starting from 1997 electronic dissemination through the Internet has become a preferred format for releasing results for many surveys. It is fair to say that some information on the sources of error and other limitations of the data should be provided to users and interpreters of the data and the amount of information provided should depend on the length, type, and importance of the report.

Users of survey data need information about a survey's quality in order to assess the survey results properly. In recognition of this need, the standards adopted by many statistical agencies specify that users should be informed of survey quality. However, the measurement of survey quality and its presentation to users are not simple tasks. There are many aspects of quality to consider, and a range of statistical indicators that may be employed to measure them. Some indicators can be produced fairly readily (e.g. response rates) but others are costly (e.g. measures of response bias). In addition, users can benefit from the availability of detailed documentation of the survey procedures since procedures used are indicative of survey quality (Kasprzyk and Kalton, 1997).

In practice the amount of information produced is often limited by budget and time constraints. The situation varies depending on whether the survey is a one-time, continuous, or periodic. Continuous and periodic surveys have the advantage that research on their methods and procedures can lead to future improvements. The systematic gathering of information about survey procedures and sources of error assists this process. Over time such programme can accumulate a substantial amount of information on data quality (Kordos, 1988).

As it has been stressed, reporting formats vary from a press release of several pages of analysis on a specific question, basic report from the survey to detailed complex analytic modelling exercises designed to test specific hypotheses. Although there is a considerable recognition of the need for reporting the extent and nature of the source of errors, there is a considerable uncertainty about the amount of detail that ought to be provided. Details reported ought to depend on the length of the report, its intended use, the nature of the data (one-time survey
vs. continuing), the survey budget, and agency policy (Depoutot, 1999; Kasprzyk et al., 1999).

The author conducted four studies to help characterise current practices of reporting sources of errors in statistical publications in Poland. The first study reviewed press release and short publications of several pages (short-format reports) from six CSO divisions. The second study reviewed basic reports of the sample surveys from the same divisions. These are printed reports resulting from a sample survey with methodology, descriptive analysis of the results, and results of the survey in tables. The third study reviewed analytic publications from some divisions. Publications that compile data from many sources and reports designed primarily to study errors sources are not included in the study. The fourth study reviewed Internet with some reports from the surveys. Handbooks and articles with description of survey methodology or data quality assessment are also mentioned.

3.1. Reporting Sources of Errors in Short-Formats Reports

A total of about 100 short-format reports was reviewed. None of the short reports discussed the study design, data quality, or survey error, none of them included a reference to technical reports. In some cases, a division news release focuses on a limited set of findings, with nothing more than a reference to the main publication or the division the data are drawn from. In other cases, the division news release includes more data and describes the study purpose, sample size and sampling error. Close to one-third (31 percent) of authors of short reports included at least one piece of information describing the purpose of the survey or analysis, the key variables or made references to sample design. Slightly less than one-fourth of these cases included the sample size (24 percent of all reports), and one-fifth described the mode of data collection. There were no references to weighting and estimation procedures. There were not significant differences in reporting some aspects of data quality in short reports between different CSO divisions.

3.2. Reporting Sources of Errors in Basic Survey Reports

In the second study, publications of basic reports from the surveys produced by 6 divisions, were reviewed. Fifty nine publications covering last five years were selected to be reviewed.

These publications were published in series "Informacje i opracowania statystyczne"("Information and Statistical Materials"). They were reviewed for their treatment of sources of sampling and nonsampling errors, data collection method, sample size, sample design and estimation procedures.

About sixty percent of the reports included sampling errors (65.2 percent for social statistics, and 51.9 percent for economic statistics). Even though nonresponse error is the most visible and well known of nonsampling error and easily obtainable, only 49.2 percent included any references to response rates, to
nonresponse as a potential source of error, or to imputations (75 percent for social statistics, and 18.5 percent for economic statistics). However, only 5 percent mentioned nonsampling errors. Measurement error was cited as a source of error in about 3.4 percent of the reports, but no mention of specific analyses of measurement bias or measurement variance were mentioned. Finally, none of the reports reported coverage rates or mentioned coverage as a potential source of error. Results are given in table 2.

Table 2. Reporting Sources of Errors and Study Design Features in Fifty Nine Sample Survey Reports

<table>
<thead>
<tr>
<th>Reporting:</th>
<th>Surveys belonging to:</th>
<th>Error sources</th>
<th>Study design features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Social statistics</td>
<td>Economic statistics</td>
<td>TOTAL</td>
</tr>
<tr>
<td></td>
<td>(demography, labour and living conditions)</td>
<td>(agriculture, enterprises, trade &amp; services)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Number</td>
<td>Percent</td>
<td>Number</td>
</tr>
<tr>
<td>Sampling</td>
<td>32</td>
<td>100.0</td>
<td>27</td>
</tr>
<tr>
<td>Nonsampling</td>
<td>2</td>
<td>6.3</td>
<td>14</td>
</tr>
<tr>
<td>Nonresponse</td>
<td>24</td>
<td>75.0</td>
<td>5</td>
</tr>
<tr>
<td>Measurement</td>
<td>1</td>
<td>3.1</td>
<td>1</td>
</tr>
<tr>
<td>Coverage</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Data collection methods</td>
<td>32</td>
<td>100.0</td>
<td>27</td>
</tr>
<tr>
<td>Sample size</td>
<td>30</td>
<td>93.7</td>
<td>15</td>
</tr>
<tr>
<td>Sample design</td>
<td>25</td>
<td>78.1</td>
<td>18</td>
</tr>
<tr>
<td>Weighting Estimation</td>
<td>20</td>
<td>62.5</td>
<td>14</td>
</tr>
</tbody>
</table>

Source: the author’s review of the CSO publications.

We have grouped divisions into two groups: survey belonging to:

- social statistics, and
- economic statistics.

It is connected with tradition and some experience in sampling. Several publications provided excellent descriptions of the sample design, data collection methods, and estimation procedures. All reports included data collection methods, and size of sample was included in 45 reports (76.3 percent) (93.7 percent for social statistics, and 55.6 percent for economic statistics); 43 reports included sample design (72.9 percent) (78.1 percent for social statistics, and 66.7 percent for economic statistics), and estimation procedures were given in 34 reports (57.6 percent) (62.5 percent for social statistics, and 51.9 percent for economic statistics).

Descriptions of the comparability of survey results with other data sources were somewhat less commonly provided, for example, 9 reports (15.3 percent) described in detail the survey changes that affect comparisons of the survey’s data over time.
3.3. Reporting Sources of Error in Analytic Publications

The study reviewed the information available on sources of error in 12 analytic publications published in series "Studia i analizy statystyczne" ("Statistical Studies and Analysis"). All publications included data collection methods, purpose of the analysis, and description of key variables. Size of sample was included in 9 publications (75 percent); sample design and estimation procedures mentioned in 3 publications (25 percent). Sampling error, the best known source of survey error and easiest to estimate, was included only in 3 publications (25 percent). Response rates are commonly used as indirect measures of nonresponse error, and are popular indicators of data quality and are expected to be routinely reported, were included in 8 publications (66.7 percent). Finally, measurement error and coverage error were not mentioned as a possible sources of nonsampling errors.

The comparison of survey data to data from independent sources was done in 9 publications (75 percent). Such comparisons can be a difficult but rewarding exercise for the data producer as well as the data user. The comparisons can result in a better understanding of the limitations of the data set for a particular form of analysis and lead to improved procedures for future rounds of the survey.

3.4. Reporting Sources of Error on the Internet

The Internet has become the medium for dissemination of data products for the CSO since May 1997. However, there are no guidelines and practices for reporting error sources over the Internet. There are some written standards for web sites, but these generally focus on web site design, layout, and administrative access. We have started to study experiences in this field of a few agencies, such as the Census Bureau, which have begun the process of developing standards for providing information about data quality over the Internet (U.S. Bureau of the Census, 1997).

This study reviewed the accessibility of data quality documentation on current Internet site of the CSO. There are printed reports prepared for press release and general utilisation by users.

The study found current Internet standards for data quality information echo the standards for printed reports and statistical tables. More explicit standards for how the advantages of the Internet media should be employed to make data quality information more easily accessible do not exist. The Internet is a dynamic medium. It is constantly changing and being upgraded. The technology and possibilities grow and change every day. We are beginners in this field, and hope to use Internet for data quality presentation efficiently.

4. Conclusions
During the period of transformation we had to change the methodology of many surveys, to apply international standards and re-organise statistics and to anticipate future needs of the users. There is still an important problem of the way in which data should be disseminated to various users, to develop methods of their adequate use, ways of interpreting and methods of statistical analysis which could be easy to understand to a wide range of users, the ways of increasing statistical literacy of the society for the informed participation and acceptance of the processes of transformation of statistics. These problems are still waiting for a solution.

The analysis is not intended to single out individual divisions. The following recommendations are offered to help divisions meet the reporting standard identified by international statistical community (e.g.: Statistics Canada, 1998; Statistics Sweden, Lyberg et al., 1998; U.K. Government Statistical Service, 1997; U.S. Census Bureau, 1998).

1. As a minimum standard, the short news releases should include the title of data collection and a mention of the fact that all report are subject to reporting errors. In addition, there should be a reference to a source report that includes more detailed information about the data collection and data quality, and a mention of the fact that sample survey data are subject to sampling errors. In addition, longer news releases and short reports of any length should include a brief description of the purpose or content of the data collection, the sample size, and the possible presence of sampling and nonsampling errors. Reports should also reference the possible sources of nonsampling errors and where appropriate draw comparisons with related data sources.

2. As it has been stressed, users of survey data need information about a survey's quality to properly assess survey results. Standards adopted by many statistical agencies specify that users should be informed of survey quality. The studies reviewed in this paper illustrate the range of division practices in reporting information on error sources in surveys. The studies suggest CSO divisions not merely define policy and standards in the reporting of such information, but monitor the implementation of this policy. In any case, data users are best served when information about survey procedures and sources of error are readily available to them to help the interpretation of the analysis. In the case of the Polish experience, more interest and emphasis in this topic is desirable.

3. Sampling error is often the only error source presented when reporting survey estimates. With main report publications, for example, it is common practice to specify sampling error without mentioning other error sources. When sampling errors are not reported, sample size may be reported, presumably as an indicator of sampling error. Sampling error may be communicated in a number of different forms, e.g., standard errors, coefficient of variation, and confidence intervals. In Polish statistical publications only relative standard
errors are presented for some estimated parameters without interpretation. For users such information is not clear.

4. **Nonresponse** is the most visible and well known source of nonsampling error. It generally reduces sample size, resulting in increased variance, and introduces a potential for bias in the survey estimates. Nonresponse rates are frequently reported and are often viewed as the first area requiring study in assessing a survey. Failure to achieve a high response rate influences perceptions of the overall quality of the survey. However, nonresponse rates provide no direct evidence on the level of nonresponse bias in survey estimates. Moreover, an examination of bias should ideally take into account any nonresponse adjustments made to attempt to reduce the bias. (Kordos, 1988).

5. **Coverage error** is the error associated with the failure to include some population units in the frame used for sample selection (undercoverage) and the error with the failure to identify units represented on the frame more than once (overcoverage). This type of error was not reported in our studies.

6. **Measurement errors** are the most difficult aspect of survey data quality to quantify. Special studies are required, and these studies are often expensive to conduct. A key distinction in categorising measurement error studies is between those that attempt to assess measurement bias and those that are concerned only with measurement variance. Studies of measurement bias need to obtain measures of "true values" with which the survey responses can be compared. Studies of measurement variance investigate only the variability of responses across different applications of the survey process, sometimes to estimate the variable error associated with a particular source (e.g., interviewer, designated respondent, or question form). Only in two surveys, the post-enumeration surveys, a measurement errors were assessed.

7. **Processing errors** include data entry, coding, and editing errors. Data entry errors may be measured though the use of a quality control sample, whereby a sample of questionnaires is selected for re-entry and an indicator of the quality of the operation, a keying error rate, is determined. Processing errors are available but not published in the reports.

8. Many statistical agencies have accepted *guidelines for survey processes*. The Polish CSO also needs to prepare such guidelines. Guidelines help an agency codify how it expects to behave professionally, help to promote consistency among studies, and document methods and principles used in collection, analysis, and dissemination. These guidelines generally include prescriptions for the dissemination of information about data quality to users. Examples include the “quality guidelines” developed by Statistics Canada (1998) and Statistics Sweden (Lyberg et al., 1998). The U.K. Government Statistical Service (1997) has developed guidelines that focus specifically on the reporting of data quality. These guidelines are in the form of a checklist of questions relating to individual areas of the survey process.
Nations (1964) presented recommendations on the topics to be documented when preparing sample survey reports, including information on many sources of error. In is important to note that information about survey procedures can also provide users with valuable insights into data quality. It is important to report information about the background and history of the data collection programme, its scope and objectives, sample design and sample size, data collection procedures, time period of data collection, the response mode, the designated respondents, as well as processing and estimation procedures. For repeated surveys, it is also important for users to be aware of changes in design, content, and implementation procedures since they may affect comparisons over time.

REFERENCES


CENTRAL STATISTICAL OFFICE (1999), Metodyka badania budżetów gospodarstw domowych (Methodology of Household Budget Survey), Zeszyty Metodologiczne, Główny Urząd Statystyczny, Warszawa, pp.146.


REMARKS TO THE UTILIZATION
OF THE CONCEPT OF LINEAR RECTIFICATIONS
OF REPETITIVE FORECASTS

Josef Kozák 1

ABSTRACT

When statistical forecasts following the analysis of a time-series are constructed, we often face situations when building an appropriate model is questionable or sometimes even practically impossible. Such situations arise mostly when the time-series are of relative small size, "good" explanatory variables are absent, etc. Therefore some acceptable solutions of the forecasting problem in these situations are of practical interest. In this respect some proposals have been published in (Kozak, 1999) in the research journal of the University of Economics in Prague. The same problems are handled also in this paper; this gives the author the possibility to repair some misunderstandings from the original paper.

Key words: repetitive forecasts, linear rectification of repetitive forecasts, average square error (ASE), linear rectified forecasts, pseudo-forecasts, active forecasts, “ex post” rectification.

1. Preliminaries

(I) Let us introduce the concept of the repetitive forecasts

\[ R_t, t = t_0, t_0 + l, \ldots, t_0 + (n - l) \]

which are to be interpreted as known "ex ante" and from the point of view of precision still acceptable forecasts of the unknown future values \( Y_t, \)

\[ t = t_0, t_0 + l, \ldots, t_0 + (n - l) \]

As an illustration, let us consider the following example. Analyzing monthly time-series, there are often situations where we have to construct forecasts of the unknown quantities \( Y_t \) for the time-point \( t_0 \) and following 11 time-points. The

1 Prof. Ing. Josef Kozák, DrSc., University of Economics, Prague, Czech Republic; e-mail: kozak@vse.cz
experience often shows that in this case the role of the repetitive forecast can be played by observations in the same months of the preceding year, i.e. the set of values

\[ R_t = Y_{t-12}, \; t = t_0, t_0 + 1, ..., t_0 + (n - 1), \; 1 < n \leq 11 \]

(II) In the given situations there is possible to construct primitive forecasts defined as suitable function of a set of values of the repetitive forecasts \( R_t, \; t = t_0 + 1, ..., t_0 + (n - 1) \). The most primitive form is linear rectification of repetitive forecasts either in the form

\[ P_t(C) = C R_t, \; t = t_0, t_0 + 1, ..., t_0 + (n - 1) \]

or in a little more complicated form

\[ P_t(a, c) = a + c R_t, \; t = t_0, t_0 + 1, ..., t_0 + (n - 1) \]

where \( C, \; a, \; c \) denote constants. These forecasts will be investigated in greater detail with a special attention to the choice of these constants.

(III) The users will widely agree that the employed forecasting procedure has to be "precise" enough (at least from the "ex post" point of view). As a measure of the "ex post" accuracy average squared error (hereinafter "ASE") can be employed, i.e. the measure

\[ ASE(P(C)) = \frac{1}{n} \sum_{t=1}^{n} (P_t(C) - Y_t)^2 \]

in the first case or the measure

\[ ASE(P(a, c)) = \frac{1}{n} \sum_{t=1}^{n} (P_t(a, c) - Y_t)^2 \]

in the second, where \( \sum \) denotes the summation over the set \( t = t_0, t_0 + 1, ..., t_0 + (n - 1) \). Under given circumstances both types of the linear rectified forecasts can be accepted, if the inequalities

\[ ASE(P(C)) \leq ASE(R) \quad \text{or} \quad ASE(P(a, c)) \leq ASE(R) \]

hold, where

\[ ASE(R) = \frac{1}{n} \sum_{t=1}^{n} (R_t - Y_t)^2 \]

denotes the ASE of the set of the pure repetitive forecasts.
(IV) Now, let us introduce some notations. Let us consider variables $U_t$,
$V_t$, etc., for $t = t_0, t_0 + 1, ..., t_0 + (n-1)$. Their mean values will be
characterized by the usual arithmetic average

$$
\overline{U} = \frac{1}{n}\sum U_t
$$

etc. In order to characterize mutual relationship of these variables a little unusual
measure - the average product

$$
M(U,V) = \frac{1}{n}\sum U_t V_t
$$

will be used. This measures comprises also the measure of variation having the
form of averaged squares

$$
M(U,U) = \frac{1}{n}\sum U_t^2
$$

The relationships between the considered variables can be also based on the usual
covariance

$$
S(U,V) = \frac{1}{n}\sum(U_t - \overline{U})(V_t - \overline{V}) = M(U,V) - \overline{U}\overline{V}
$$

and variance

$$
S(U,U) = \frac{1}{n}\sum(U_t - \overline{U})^2 = M(U,U) - \overline{U}^2
$$

(V) For $t = t_0, t_0 + 1, ..., t_0 + (n-1)$ let us consider a set of forecasts $P_t$ of
predicted quantities $Y_t$ leading to the ASE

$$
ASE(P) = \frac{1}{n}\sum(P_t - Y_t)^2
$$

This quantity can be expressed as

$$
ASE(P) = M(P,P) - 2M(P,Y) + M(Y,Y)
$$

or in a little more complicated form as

$$
ASE(P) = (S(P,P) - 2S(P,Y) + S(Y,Y)) + (\overline{P} - \overline{Y})^2
$$

(VI) As far I know, the construction of linear rectifications of repetitive
forecasts employed for the solution of the problem of prediction has not yet been
considered by other authors. Therefore the goal of further remarks is to consider
some possibilities of the selection of parameters \( C, a, c \) in two special situations.

(a) Initially this selection will be considered in the situation of the analysis of so-
called pseudo-forecasts where values of \( Y, R \),
\[ t = t_0, t_0 + 1, ..., t_0 + (n - 1), \]
are known.

(b) Then the selection of constants will be considered in the more practical
situation of the so-called active forecasts, i.e. when future values \( Y \),
\[ t = t_0, t_0 + 1, ..., t_0 + (n - 1), \]
are not known.

"EX POST" RECTIFICATION

Let us consider the choice of \( C, a, c \) in linear rectifications of repetitive
forecasts, which is feasible in the situation that both sets \( Y, R \),
\[ t = t_0 + 1, ..., t_0 + (n - 1), \]
are known. This kind of analysis will be called the analysis of accuracy of pseudo-forecasts.

(I) First, let us consider the simple linear rectification.

(I-1) The initial problem is the choice of \( C \) leading to the minimum value of
the ASE. Obviously
\[
ASE(P(C)) = C^2 \, M(R, R) - 2 \, C \, M(R, Y) + M(Y, Y)
\]
is a quadratic function of \( C \) with the first derivative
\[ 2 \, C \, M(R, R) - 2M(R, Y) \]
Therefore for \( C = B \) with
\[
B = \frac{M(R, Y)}{M(R, R)}
\]
ASE leads to the minimum value equal to
\[
ASE(P(B)) = M(Y, Y) - \frac{M^2(R, Y)}{M(R, R)}
\]

(I-2) Despite the assertion about the minimum of the ASE(P(B)) the
following fact has to be kept in mind - the pure repetitive forecasts represent the
simplest and from the viewpoint of the accuracy the limiting solution of the
problem of prediction. It seems to be useful to know how the considered version
of linear rectification will hold in the ASE in comparison with the set of pure repetitive forecasts. This approach leads to the analysis of the difference of both ASE’s

\[ D(C) = ASE(R) - ASE(P(C)) \]

which can be expressed as

\[ D(C) = (1 - C^2) M(R, R) - 2(1 - C) M(R, Y) \]

and with regard to the definition of \( B \) also as

\[ D(C) = F(C) M(R, R) \]

with

\[ F(C) = (1 - C) (1 + C - 2B) = (1 - B)^2 - (C - B)^2 \]

Thus the required inequality

\[ ASE(P(C)) \leq ASE(R) \]

can be achieved with \( C \)’s leading to \( F(C) \geq 0 \). The required situation takes place for

\[ (1 - C) \geq 0 \text{ together with } (1 + C - 2B) \geq 0 \]

i.e. for \( C \)’s satisfying condition

\[ 2B - 1 \leq C \leq 1 \]

further, the required inequalities are valid also for

\[ (1 - C) \leq 0 \text{ together with } (1 + C - 2B) \leq 0 \]

i.e. for \( C \)’s in the interval

\[ 1 \leq C \leq 2B - 1 \]

Therefore, the following choices of \( C \) can be mentioned.

1. The medium option \( P_t(B) = B R_t, t = t_0, t_0 + 1, ..., t_0 + (n - 1) \) can be recommended. As \( F(B) = (1 - B)^2 \) is the maximum positive value of the function \( F(C) \), the recommended set leads to the minimum value of the ASE(P(B)), i.e. is the best one in ASE when compared to the pure repetitive forecasts.

2. Further, two boundary options of the given type of linear rectifications with

\[ C = 1 \text{ and } C = (2B - 1) \]
can be recommended, i.e. the sets
\[ P_1(1) = R_1 \quad \text{and} \quad P_1(2B - 1) = (2B - 1) R_1 \quad \text{for} \quad t = t_0, t_0 + 1, \ldots, t_0 + (n - 1) \]
both with the property
\[ F(1) = F(2B - 1) = 0 \]
i.e. with their ASE's equal to the ASE of the set of pure repetitive forecasts.

(II) Now, let us analyze the more complicated linear rectification.

(II-1) Introducing \( a = \overline{Y} - c \overline{R} \), we get
\[ \text{ASE}(P(a, c)) = c^2 S(R, R) - 2c S(R, Y) + S(Y, Y) \]
what is a quadratic function of the argument \( c \) with the first derivative equal to
\[ 2c S(R, R) - 2S(R, Y) \]
Thus choosing \( c = b \) with
\[ b = \frac{S(R, Y)}{S(R, R)} \]
we achieve the considered type of linear rectification with minimum value of ASE
\[ \text{ASE}(P(a, b)) = S(Y, Y) - \frac{S^2(R, Y)}{S(R, R)} \]

(II-2) Then the selection of \( c \) can be based on the analysis of the difference of ASE's
\[ D(a, c) = \text{ASE}(R) - \text{ASE}(P(a, c)) \]
Using the identity
\[ \text{ASE}(R) = (S(R, R) - 2S(R, Y) + S(Y, Y)) + (\overline{R} - \overline{Y})^2 \]
the above introduced quantity can be expressed as
\[ D(a, c) = (\overline{R} - \overline{Y})^2 + f(c) S(R, R) \]
where
\[ f(c) = (1 - c)(1 + c - 2b) = (1 - b)^2 - (c - b)^2 \]
is the quadratic function of the parameter \( c \). Thus the quantity \( D(a, c) \) comprises - compared with \( D(C) \) - the non-negative term \( (\overline{R} - \overline{Y})^2 \), which can be interpreted as the lower limit of the difference of the ASE for the given type of
linear rectifications as compared with the ASE of the pure repetitive forecasts. Further, in analogy to the one-parameter linear rectification it seems to be recommendable the construction of the medium option of linear rectifications of the form

\[ P_t(a, b) = \bar{Y} + b (R_t - \bar{R}), ~ t = t_0, t_0 + 1, ..., t_0 + (n - 1) \]

leading to the maximum \( f(b) = (1 - b)^2 \). Further, two boundary options with \( c = 1 \) and \( c = (2b - 1) \) can also be considered in the forms

\[ P_t(a, 1) = \bar{Y} + (R_t - \bar{R}) \quad \text{and} \quad P_t(a, 2b - 1) = \bar{Y} + (2b - 1) (R_t - \bar{R}) \]

for \( t = t_0, t_0 + 1, ..., t_0 + (n - 1) \). Because in these cases \( f(1) = f(2b - 1) = 0 \), the considered difference of ASE is reduced to the quantity \( (\bar{R} - \bar{Y})^2 \).

**EX ANTE" RECTIFICATION**

The preceding considerations have focused on the analysis of pseudo-forecasts. Now we will be interested on getting "ex ante" estimates of the future unknown quantities \( Y_t \) for \( t = t_0, t_0 + 1, ..., t_0 + (n - 1) \), where \( t_0 \) represents the starting point of forecasts. These estimates have to be "good enough" in the sense that their ASE must not be worse in comparison with the ASE(R) of the known pure repetitive forecasts. The crucial obstacle in construction of such forecasts is the fact, that in the given situations the constants \( B, c \) are not known. Let us emphasize the most important facts connected with the requirement to reach the inequalities \( D(C) \geq 0 \) with unknown \( B \) and \( D(a, c) \geq 0 \) for unknown \( b \) and \( \bar{Y} \).

(I) Let us start with the first type of linear rectification.

(a) First, the trivial choice \( C = 1 \) leading to \( F(1) = 0 \), i.e. to the set of "pure" repetitive forecasts, has to be mentioned.

(b) Let us now continue with prior information regarding the unknown \( B \) in form of inequalities

\[ B_0 \leq B \text{ with known } B_0 > 0 \]

Due to the fact that in this situation the relation

\[ F(C) \leq (1 - C) (1 + C - 2B_0) \]

holds, the requirement \( D(C) \geq 0 \) is valid for \( C \)'s fulfilling inequality

\[ (1 - C) (1 + C - 2B_0) \geq 0 \]
This requirement can be satisfied in two cases: in the case $0 < B_0 < 1$ for the triple of Cs

$$2B_0 - 1 < B_0 < 1$$

and in the other case $B_0 > 1$ for the triple of Cs

$$1 < B_0 < 2B_0 - 1$$

(II) Regarding the more complicated type of linear rectifications, the main obstacle is the fact that $b$ and $\bar{Y}$ are now unknown.

(II-a) First, let us limit ourselves on the solution for the most frequent situation with

$$Y_i > 0, \ R_i > 0, \ t = t_0, t_0 + 1, ..., t_0 + (n - 1)$$

under prior information

$$0 < \bar{Y} \leq q_0 \ R$$

where $q_0 > 0$ denotes a known constant. Then instead of the unreliable choice of

$$a = \bar{Y} - c \ R \leq a_0$$

the following choice

$$a_0 = (q_0 - c) \ R$$

can be considered in this case. This means that instead of the above mentioned unreliable linear rectification the form

$$P_i(a, c) = \bar{Y} + c (R_i - \bar{R}), \ t = t_0, t_0 + 1, ..., t_0 + (n - 1)$$

or another form

$$P_i(a_0, c) = q_0 \ R + c (R_i - \bar{R}), \ t = t_0, t_0 + 1, ..., t_0 + (n - 1)$$

will be used. The properties of this set can be described in the following manner.

We can write

$$P_i(a_0, c) - Y_i = c (R_i - \bar{R}) - (Y_i - \bar{Y}) - (\bar{Y} - q_0 \ R)$$

$$t = t_0, t_0 + 1, ..., t_0 + (n - 1)$$

thus, utilizing the above introduced definitions, the ASE of the considered set can be expressed as

$$ASE(P(a_0, c)) = \left(c^2 \ S(R, R) - 2c \ S(R, Y) + S(Y, Y)\right) + \left(\bar{Y} - q_0 \ R\right)^2$$
Using above mentioned decomposition of \( \text{ASE}(R) \) the difference
\[
D(a_0, c) = \text{ASE}(R) - \text{ASE}(P(a_0, c))
\]
can be expressed as
\[
D(a_0, c) = H(q_0) + f(c) S(R, R)
\]
where as previously
\[
f(c) = (1 - c) (1 + c - 2b) = (1 - b)^2 - (c - b)^2
\]
denotes a quadratic function of \( c \) and
\[
H(q_0) = \left( \bar{Y} - \bar{R} \right)^2 - \left( \bar{Y} - q_0 \bar{R} \right)^2
\]
is a quadratic function of \( q_0 \). This quantity can be expressed as
\[
H(q_0) = (1 - q_0^2) \bar{R}^2 - 2(1 - q_0) \bar{Y} \bar{R}
\]
If above mentioned inequality between averages \( \bar{Y} \) and \( \bar{R} \) are fulfilled, then inequalities
\[
H(q_0) \geq (1 - q_0^2) \bar{R}^2 \geq 0
\]
hold, which means that the first element on the right side of the analyzed difference \( D(a_0, c) \) is positive.

(II-b) Further, let us assume that at least prior information of the inequalities
\[
b_0 \leq b \quad \text{with known} \quad b_0 > 0
\]
is at our disposal "ex ante". In this situation it will be the sufficient to choose the parameter \( c \) fulfilling inequalities
\[
(1 - c) (1 + c - 2b_0) \geq 0
\]
Thus, in the situation of \( 0 < b_0 < 1 \), the triple of \( cs \) is related to inequality
\[
2b_0 - 1 < b_0 < 1
\]
or, in the other situation \( b_0 > 1 \), the triple of \( cs \) is related to inequalities
\[
1 < b_0 < 2b_0 - 1
\]
which are acceptable for our purposes - both due to the fact, that they satisfy the condition \( f(c) \geq 0 \).
Acknowledgement

I must to express my gratitude to the referee of this paper for stimulating suggestions as well as to my friend Charles Freedom for the technical preparation of the manuscript.

REFERENCES


ESTIMATION OF FINITE POPULATION DISTRIBUTION FUNCTION USING MULTIVARIATE AUXILIARY INFORMATION

M S Ahmed¹ and Walid Abu-Dayyeh²

ABSTRACT

This paper derives the generalized estimator for the finite population distribution function using multivariate auxiliary information. The properties of this estimator are given for the general sampling design. The results for SRSWOR are presented for the estimator. Finally, a simulation study is carried out from a real data set for the relative comparisons with Rueada and Arcos (1998).

Key Words: Finite population, multivariate auxiliary information, distribution function.

1. Introduction

Chambers and Dunstan (1986) and Rao et al. (1990) suggested some estimators for estimating the finite population quantiles by reversing improved estimates of the distribution functions using an auxiliary variable. Rueada and Arcos (1998) proposed an estimator by using multivariate auxiliary information for estimating the distribution function. In this paper, we propose a generalized estimator of distribution function by using multivariate auxiliary information.

Suppose \( U = (1, 2, \ldots, N) \) is a finite population of size \( N \) and any random sample, \( s \), of size \( n \), is drawn according to a sampling plan, \( \phi(s) \), with inclusion probabilities \( \pi_i \).

The population cumulative distribution function of any random variable \( V \) is

\[
F_{\nu}(v_j) = N^{-1} \sum_{i \in \tau} \Delta(v_i - v_j) \tag{1.1}
\]

where \( \Delta(v_i - v_j) = 1 \), when \( v_j \leq v_i \) and \( \Delta(v_i - v_j) = 0 \) otherwise.

¹ On leave from Department of Statistics, Jahangirnagar University, Bangladesh.
The customary design based unbiased estimator of $F_v$ is

$$\hat{F}_v(v_\tau) = \sum_{i \in s} \pi_i^{-1} \Delta(v_\tau - v_i) / \sum_{i \in s} \pi_i^{-1}$$

(1.2)

then $E(\hat{F}_v(v_\tau)) = F_v(v_\tau) = \tau$

This is true for any random variable $V$. 

\[ \text{Department of Statistics, Yarmouk University, Irbid, Jordan; e-mail: m_s_ahmed@yahoo.com} \]
2. Generalized estimator and its properties

Suppose \( Y \) and \( X_j \) (\( j=1,2,\ldots,k \)) are the survey variable and \( k \)-auxiliary variables, respectively. The purpose is to find the confidence interval, \( y_\tau \) for any given \( \tau \) by using \( k \)-auxiliary variables. We observe \((x_i,y_i)\) for \( i \in s \) and assume that finite population distribution function of \( X_j \)’s (\( j=1,2,\ldots,k \)) are known.

If a single auxiliary variable \( X \) is correlated with \( Y \), then ratio estimator

\[
\hat{F}_R (y_\tau) = \frac{\hat{F}_Y (y_\tau)}{\hat{F}_X (x_\tau)} = \frac{\hat{F}_Y (y_\tau)}{\hat{F}_X (x_\tau)} \tau
\]  
(2.1)

Rueda and Acros (1998) considered a multivariate ratio type estimator of \( \hat{F}_Y (y_\tau) \) by using \( k \)-auxiliary variables \( X_j \) (\( j=1,2,\ldots,k \))

\[
\hat{F}_{mr} (y_\tau) = \sum_{j=1}^{k} \omega_j (\hat{F}_Y (y_\tau) / \hat{F}_X (x_\tau)) \hat{F}_{X_j} (x_\tau) = \sum_{j=1}^{k} \omega_j \hat{F}_{R_j} (y_\tau) = \omega' \hat{F}_R
\]  
(2.2)

where \( \hat{F}_R = (\hat{F}_{R_1}, \hat{F}_{R_2}, \ldots, \hat{F}_{R_k})' \) and \( \omega = (\omega_1, \omega_2, \ldots, \omega_k)' \), such that \( e' \omega = 1 \), \( e = (1 1 1 \ldots 1)' \).

Now, the variance of \( \hat{F}_{mr} (y_\tau) \) is given by

\[
V(\hat{F}_{mr} (y_\tau)) = \omega' \Lambda \omega
\]  
(2.3)

where \( \Lambda = (\Lambda_{jj'}) \) is the dispersion matrix with \( \Lambda_{jj'} = Cov(\hat{F}_{R_j}, \hat{F}_{R_{j'}}) \) for \( j, j' = 1,2,\ldots,k \)

The optimum choices of \( \omega \) is \( \omega_{opt} = \Lambda^{-1} e / (e' \Lambda^{-1} e) \) and the minimum variance is

\[
V(\hat{F}_{mr} (y_\tau))_{min} = (e' \Lambda^{-1} e)^{-1}
\]  
(2.4)

We propose a generalized estimator of \( F_Y (y_\tau) \)

\[
\hat{F}_{mw} (y_\tau) = \sum_{j=1}^{k} \theta_j \hat{F}_{R_j} (y_\tau) + \theta_0 \hat{F}_Y (y_\tau)
\]  
(2.5)

such that \( \sum_{j=0}^{k} \theta_j = 1 \)

Note that if \( \theta_0 = 0 \), it is the same as (2.2) and if \( \theta_j = 0 \) for all \( j=1,2,\ldots,k \) then it is simply (2.1)

Rewrite, \( \hat{F}_{mw} (y_\tau) = \hat{F}_Y (y_\tau) - \sum_{j=1}^{k} \theta_j (\hat{F}_Y (y_\tau) - \hat{F}_{R_j} (y_\tau)) = t_0 - \theta' t \)  
(2.6)
where $\hat{F}_y(y_1) = t_0$, $\hat{F}_y(y_1) - \hat{F}_y(y_k) = t_j$, $t = (t_1, t_2, ..., t_k)'$ and 
$\theta = (\theta_1, \theta_2, ..., \theta_k)'$

Now, the variance of $\hat{F}_{mw}(y_1)$ is

$$V(\hat{F}_{mw}(y_1)) = \sigma^2_{t_0} - 2m^1\theta + \theta'M\theta$$  \hspace{1cm} (2.7)

where $E(t^2_0) = \sigma^2_{t_0}$, $M = (M_{jj'})$ is the dispersion matrix with $M_{jj'} = Cov(t_j, t_{j'})$ and $m = (m_j)$ with $m_j = Cov(t_j, t_0)$ for $j, j' = 1, 2, ..., k$.

Thus minimizing $V(\hat{F}_{mw}(y_1))$ is equivalent to finding the best homogeneous linear predictor of $t_0$ in terms of $t_1, t_2, ..., t_k$. Then (2.7) shows that $V(\hat{F}_{mw}(y_1))$ is quadratic and convex. The optimal choice of $\theta$ which minimizes $V(\hat{F}_{mw}(y_1))$ is easily obtained as $\theta = M^{-1}m$ and minimum value of $V(\hat{F}_{mw}(y_1))$ is

$$V_{min}(\hat{F}_{mw}(y_1)) = \sigma^2_{t_0} - m'M^{-1}m$$  \hspace{1cm} (2.8)

3. Results for simple random sampling with two auxiliary variables

Suppose a simple random (without replacement) sample of size $n$ is drawn from a finite population of size $N$, then the unbiased estimate of $F_v(v_1)$ is

$$\hat{F}_v(v_1) = \sum_{i=0}^{n-1} \Delta(v_i - v_1) / n$$ \hspace{1cm} (3.1)

Then $E(\hat{F}_v(v_1)) = \tau$ and $V(\hat{F}_v(v_1)) = (n^{-1} - N^{-1})\tau(1 - \tau)$ for $v = y$ and $x_j$

$j = 1, 2, ..., k$

Suppose $P_{jj'}$ is the proportion of population units with $x_j \leq x_{j'}$ and $x_{j'} \leq x_j$, then $P_{jj'} = \tau$ if $j = j'$.

Now we have,

$M_{jj'} = Cov(t_j, t_{j'}) = Cov(\hat{F}_{x_j}(x_{j'}), \hat{F}_{x_j}(x_{j'})) = (n^{-1} - N^{-1})(P_{jj'} - \tau^2)$,

$m_{jj'} = Cov(t_j, t_0) = Cov(\hat{F}_{x_j}(x_j), \hat{F}_y(y_1)) = (n^{-1} - N^{-1})(P_{jj'} - \tau^2)$,
\[ \Lambda_{jj'} = \text{Cov}(\hat{F}_{j}, \hat{F}_{j'}) \]
\[ \cong \text{Var}(\hat{F}_{j}(y_{j})) - \text{Cov}(\hat{F}_{j}(y_{j}), (\hat{F}_{j}(x_{j}) + \hat{F}_{j'}(x_{j'}))) + \text{Cov}(\hat{F}_{j}(x_{j}), \hat{F}_{j'}(x_{j'})) \]
\[ \Lambda_{jj'} = \text{Cov}(\hat{F}_{j}, \hat{F}_{j'}) \cong (n^{-1} - N^{-1}) (\tau - P_{yj} - P_{yj'} + P_{yj}) \]
\[ \Lambda_{jj} = V(\hat{F}_{j}) \cong 2(n^{-1} - N^{-1}) (\tau - P_{yj}) , \]
\[ \rho_{jj'} = \text{Corr}(\hat{F}_{j}(x_{j}), \hat{F}_{j'}(x_{j'})) = (P_{yj'} - \tau^2) / \tau (1 - \tau) \]
\[ \rho_{yj} = \text{Corr}(\hat{F}_{j}(y_{j}), \hat{F}_{j}(x_{j})) = (P_{yj} - \tau^2) / \tau (1 - \tau) \]
\[ M_{jj'} = (n^{-1} - N^{-1}) \tau (1 - \tau) \rho_{yj}, m_{yj} = (n^{-1} - N^{-1}) \tau (1 - \tau) \rho_{yj} \]
\[ \Lambda_{jj'} \cong (n^{-1} - N^{-1}) \tau (1 - \tau) (1 - \rho_{yj} - \rho_{yj'} + \rho_{yj'}) \]
\[ \Lambda_{jj} = V(\hat{F}_{j}) \cong 2(n^{-1} - N^{-1})(1 - \rho_{yj}) \]

Now, the detailed results are given for two auxiliary variables. For k=2, we have
\[ \Lambda = (n^{-1} - N^{-1}) \tau (1 - \tau) \left( \begin{array}{ccc}
2(1 - \rho_{y1}) & 1 - \rho_{y1} - \rho_{y2} + \rho_{12} & 1 - \rho_{y1} - \rho_{y2} + \rho_{12} \\
1 - \rho_{y1} - \rho_{y2} + \rho_{12} & 2(1 - \rho_{y2}) & 1 - \rho_{y1} - \rho_{y2} + \rho_{12} \\
1 - \rho_{y1} - \rho_{y2} + \rho_{12} & 1 - \rho_{y1} - \rho_{y2} + \rho_{12} & 2(1 - \rho_{y2}) \end{array} \right) \]
\[ \omega_{opt} = \Lambda^{-1} e / (e' \Lambda^{-1} e) = \omega_{opt} = \left( \frac{\omega_{01}}{\omega_{02}} \right) = \frac{1}{2(1 - \rho_{12})} \left( \frac{1 + \rho_{y1} - \rho_{y2} - \rho_{12}}{1 - \rho_{y1} + \rho_{y2} - \rho_{12}} \right) \]
\[ V(\hat{F}_{kn}(y_{j}))_{min} = (n^{-1} - N^{-1}) \tau (1 - \tau) \frac{4(1 - \rho_{y1})(1 - \rho_{y2}) - (1 - \rho_{y1} - \rho_{y2} + \rho_{12})^2}{2(1 - \rho_{12})} \quad (3.2) \]
\[ \theta = M^{-1} m = \frac{1}{1 - \rho_{12}^2} \left( \rho_{y1} - \rho_{y2} \rho_{12} \right) \quad \sigma_{\theta}^2 = (n^{-1} - N^{-1}) \tau (1 - \tau) \text{ and} \]
\[ m' M^{-1} m / \sigma_{\theta}^2 = (1 - \rho_{12}^2)^{-1} (\rho_{y1}^2 + \rho_{y2}^2 - 2 \rho_{y1} \rho_{y2} \rho_{12}) \]
\[ V_{min}(\hat{F}_{mn}(y_{j})) = (n^{-1} - N^{-1}) \tau (1 - \tau) [1 - (1 - \rho_{12}^2)^{-1} (\rho_{y1}^2 + \rho_{y2}^2 - 2 \rho_{y1} \rho_{y2} \rho_{12})] \quad (3.3) \]

The estimates of these may be obtained by replacing \( \rho_{yj} \) by its estimate
\[ \hat{\rho}_{yj} = \tau^{-1}(1 - \tau)^{-1}(p_{yj} - \tau^2), \text{ where } p_{yj} \text{ is the proportion units with } y \leq y_{k} \text{ and } x_{j} \leq x_{j'}, \text{ in the sample.} \]
In the following section we give a simulation study with a real data to compare the above results.

4. Numerical Illustrations with simulation and conclusions

The relative comparisons among these methods are made by using a real data set. The data for the illustration has been taken from Koi thasil, District Handbook of Aligarh, India, 1990, of 340 villages. We consider $y$ to be the number of cultivators, $x_1$ is the area, and $x_2$ is the number of households in a village. A simulation study of 10,000 repeated samples without replacement of sizes 50 and 70 (15% and 20%) respectively, for $F_y = 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80$ and 0.90. We compute the expected value, variance and percentage relative gain in efficiency (PRGE) over simple method for 10,000 repeated samples for different estimators. The percentage relative gain in efficiency of $\hat{F}_{mr}$ and $\hat{F}_{mw}$ over $\hat{F}_y$ are defined respectively by

$$\text{PRGE}(\hat{F}_{mr}) = \frac{V_m(\hat{F}_y) - V_m(\hat{F}_{mr})}{V_m(\hat{F}_{mr})} \times 100$$

and

$$\text{PRGE}(\hat{F}_{mw}) = \frac{V_m(\hat{F}_y) - V_m(\hat{F}_{mw})}{V_m(\hat{F}_{mw})} \times 100$$

<table>
<thead>
<tr>
<th>$F_y$</th>
<th>$n$</th>
<th>$\hat{F}_y$</th>
<th>$\hat{F}_{mr}$</th>
<th>$\hat{F}_{mw}$</th>
<th>$\hat{F}_y$</th>
<th>$\hat{F}_{mr}$</th>
<th>$\hat{F}_{mw}$</th>
<th>$\text{PRGE}_{mr}$</th>
<th>$\text{PRGE}_{mw}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>50</td>
<td>0.10296</td>
<td>0.10487</td>
<td>0.10554</td>
<td>0.00180</td>
<td>0.00117</td>
<td>0.00101</td>
<td>54.44</td>
<td>77.53</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.10277</td>
<td>0.10628</td>
<td>0.10600</td>
<td>0.00129</td>
<td>0.00080</td>
<td>0.0007</td>
<td>60.55</td>
<td>82.57</td>
</tr>
<tr>
<td>0.2</td>
<td>50</td>
<td>0.20180</td>
<td>0.20844</td>
<td>0.20668</td>
<td>0.00320</td>
<td>0.00176</td>
<td>0.00156</td>
<td>81.99</td>
<td>104.61</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.20243</td>
<td>0.20811</td>
<td>0.20740</td>
<td>0.00229</td>
<td>0.00115</td>
<td>0.00103</td>
<td>99.19</td>
<td>121.37</td>
</tr>
<tr>
<td>0.3</td>
<td>50</td>
<td>0.30200</td>
<td>0.30341</td>
<td>0.30330</td>
<td>0.00420</td>
<td>0.00210</td>
<td>0.00191</td>
<td>99.94</td>
<td>120.04</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>0.30263</td>
<td>0.30197</td>
<td>0.30191</td>
<td>0.00300</td>
<td>0.00140</td>
<td>0.00128</td>
<td>114.6</td>
<td>134.48</td>
</tr>
<tr>
<td>0.4</td>
<td>50</td>
<td>0.40300</td>
<td>0.40428</td>
<td>0.40473</td>
<td>0.00480</td>
<td>0.00234</td>
<td>0.00213</td>
<td>104.8</td>
<td>125.25</td>
</tr>
<tr>
<td>( \theta )</td>
<td>0.40231</td>
<td>0.39972</td>
<td>0.40079</td>
<td>0.00334</td>
<td>0.00158</td>
<td>0.00144</td>
<td>0.00147</td>
<td>0.00369</td>
<td>0.00132</td>
</tr>
<tr>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>0.5</td>
<td>0.50280</td>
<td>0.50521</td>
<td>0.50504</td>
<td>0.00500</td>
<td>0.00277</td>
<td>0.00247</td>
<td>0.00172</td>
<td>0.00198</td>
<td>0.00123</td>
</tr>
<tr>
<td>0.6</td>
<td>0.60168</td>
<td>0.60295</td>
<td>0.60207</td>
<td>0.00480</td>
<td>0.00265</td>
<td>0.00236</td>
<td>0.00174</td>
<td>0.00165</td>
<td>0.00108</td>
</tr>
<tr>
<td>0.7</td>
<td>0.70184</td>
<td>0.69941</td>
<td>0.69804</td>
<td>0.00420</td>
<td>0.00207</td>
<td>0.00187</td>
<td>0.00143</td>
<td>0.00126</td>
<td>0.00080</td>
</tr>
<tr>
<td>0.8</td>
<td>0.80136</td>
<td>0.79977</td>
<td>0.79922</td>
<td>0.00320</td>
<td>0.00152</td>
<td>0.00135</td>
<td>0.00098</td>
<td>0.00084</td>
<td>0.00055</td>
</tr>
<tr>
<td>0.9</td>
<td>0.90831</td>
<td>0.90225</td>
<td>0.90388</td>
<td>0.00138</td>
<td>0.00069</td>
<td>0.00059</td>
<td>0.00034</td>
<td>0.00032</td>
<td>0.00020</td>
</tr>
</tbody>
</table>

It is observed that the estimators \( \hat{F}_Y \), \( \hat{F}_m \) and \( \hat{F}_m \) are all almost unbiased for 10,000 repeated samples under simulation. From the percentage relative gain in efficiency of \( \hat{F}_m \) and \( \hat{F}_m \) over \( \hat{F}_Y \) it is observed that \( \hat{F}_m \) gains more than \( \hat{F}_m \) over \( \hat{F}_Y \) for cases. The gain of \( \hat{F}_m \) increases slightly more for the large value of \( F_Y \).

**REFERENCES**


REPORT

The Tenth Didactical Conference on Methods of Qualitative Attributes Analysis in Decision Making Process
University of Lodz, Poland, 4 - June 2001

The Tenth Didactical Conference on Methods of Qualitative Attributes Analysis in Decision Making Process, was held in Lodz, Poland, from 4 to 5 June 2001. The Conference was organised by the Institute of Econometrics and Statistics, of the University of Lodz. Didactical conferences are a discussion forum for Polish mathematicians, statisticians, econometricians and information specialists focused on finding good teaching programs for quantitative subjects, taught on an economic, information and management and marketing courses. This year’s Conference was primarily devoted to the methods of analysing Decision Making Process with special attention to qualitative attributes. These methods belong to advanced quantitative analyses, rarely found in the teaching programs of such subjects as: mathematics, statistics, econometrics and informatics systems, especially after limiting the number of teaching hours by the so called program minima. At four sessions 15 papers were presented as well as some applications of STATISTICA package to the analyses of qualitative data, presented by the StatSoft staff. Among 110 Conference participants, beside university lecturers, were the representatives of Municipal Authorities of Lodz, banks and Statistics Office. The honorary quest was Prof. Adam Biela – member of the Polish Parliament, who presented the opening paper, entitled Multivariate scaling in social and economic research. The Chairman of the Conference Program Committee was Prof. Czesław Domański and the Chairman of the Conference Organisation Committee was Prof. Jadwiga Suchecka.

At the data analysis in Decision Making Process session the following papers were presented:

• Miroslaw Szreder – Bayesian Methods in Making Decisions by Manufacturers, in which he presented the application of Bayesian decision theory to a typical marketing problem;

• Czesław Domański – Information in Decision Making Process, discussing basic measurement scales and their role in creating information necessary for future decision makers; Marek Michalak – Activities of Founding Organ as Supervising Companies in Privatization Processes;
• Marek Melaniuk – Benchmarking as a Method of Upgrading Decision Making Process, in which he discussed problems connected with the application of benchmarking in decision making process.

At the quantitative data in teaching quantitative session the following papers were presented:
• Krystyna Pruska – Probit and Logit Models in Economic Studies Programs,
• Lechoslaw Stepien – Application of Correlation Measures for Qualitative Attributes in Decision Making Process,
• Artur Gajdos and Danuta Rozpedowska-Ptycia – Computer Packages for Modelling Qualitative.

At the qualitative attributes in scientific research session the following papers were presented:
• Stanislaw Krawczyk – Methods of Classifying Objects with Respect to Qualitative Attributes, in which he overviewed and compared classification methods for objects defined through qualitative attributes;
• Eugeniusz Gatnar – Qualitative Attributes in Discriminate Analysis, in which he presented a discrimination model for an exemplary set of qualitative data;
• Agnieszka Rossa – Graphical Log-linear Models and their Applications in Multivariate Analysis of Qualitative Data;
• Cyprian Kozyra – Methods of Qualitative Data Analysis.

At the last data session on changes in the organisation of the teaching of quantitative subjects on economic courses the following papers were presented:
• Lucja Tomaszewicz, Jakub Boratynski – Potential Directions of Changes in the Organisation of Universities Education System,
• Halina Klepacz, Elzbieta Zoltowska – The Quality of Mathematical Knowledge of Freshmen of Economic Studies,
• Elzbieta Zoltowska, Halina Klepacz, Elzbieta Porazinska – New Maturity Exam in Mathematics and Realisation of Current Programs at Economic Studies.


Anna Szymanska
Announcement

8th CONFERENCE OF THE INTERNATIONAL FEDERATION OF CLASSIFICATION SOCIETIES
Cracow, Poland, July 16-19, 2002

The Conference will take place under the auspices of the International Federation of Classification Societies (IFCS) and is organised by Sekcja Klasyfikacji i Analizy Danych PTS (Section for Classification and Data Analysis of the Polish Statistical Society). Previous conferences were held in Aachen (Germany,1987), Charlottesville (USA,1989), Edinburgh (UK,1991), Paris (France,1993), Kobe (Japan,1996), Rome (Italy,1998) and Namur (Belgium, 2000).

The member societies participating in the IFCS are the British Classification Society, Associacao Portugesa de Classificapao e Analise de Dados, Classification Society of North America, Gesellschaft fur Klassifikation, Japanese Classification Society, Korean Classification Society, Societe Francophone de Classification, Societa Italiana di Statistica, Sekcja Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, Vereniging voor Ordinatie en Classificatie, Irish Pattern Recognition and Classification Society, Central American and Carribean Society of Classification and Data Analysis.

The IFCS is a non-profit and non-political scientific organisation which promotes the dissemination of technical and scientific information concerning data analysis, classification, related methods, and their applications. The IFCS encourages young researchers to attend its conferences, and special arrangements will be offered to them for the IFCS-2002 in Cracow.

SCIENTIFIC PROGRAM COMMITTEE
Chair
K. Jajuga (Poland)

International Members
P. Arabie (USA), D. Banks (USA), H. Bacelar Nicolau (Portugal), H. H. Bock (Germany), P de Boeck (Belgium), F da Costa Nicolau (Portugal), A. Ferligoj (Slovenia), B. Fichet (France), W. Gaul (Germany), A. Gordon (UK), J. Gower (UK), P Groenen (The Netherlands), C. Hayashi (Japan), S. Hirtle (USA), H. Kiers (The Netherlands), C. Lauro (Italy), L. Lebart (France), S. Lee (Korea),
F. Murtagh (Ireland), N. Ohsumi (Japan), O. Opitz (Germany), J.-P. Rasson (Belgium), G. Ross (UK), J. Trejos (Costa Rica), M. Vichi (Italy)

Members from the Polish Classification Society (SKAD)

LOCAL ORGANIZING COMMITTEE
Chair
A. Sokołowski

CONFERENCE ADDRESSES
Local Organizing Committee IFCS·2002
Andrzej Sokołowski, Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland, phone: +48 12 2935003, fax: +48 12 2935004, e-mail: ifcs2002@ae.krakow.pl

Scientific Program Committee IFCS·2002
Krzysztof Jajuga, Wrocław University of Economics, ul. Komandorska 118/120, 53-345 Wrocław, Poland
phone: +48 71 3680340, +48 71 3680338, fax: +48 71 3680322, e-mail: ifcs2002@credit.ae.wroc.pl

IFCS-2002 Bank Account
BPH IV O/Krakow, swift code: BPHKPLPK, Account No.10601389-380000049516
Akademia Ekonomiczna IFCS-2002

World Wide Web
http://ifcs2002.ae.krakow.pl

General Topics
- Probabilistic Methods - Validation
- Bayesian Data Analysis - Computational Methods
- Data Mining - Combinatorial Algorithms
- Mixture Models - Genetic Algorithms
- Internet Survey - Information Retrieval
- Graphs - Neural Networks

Classification
- Decision Trees - Hierarchical Methods
- Phylogenetic Methods - Discriminant Analysis
- Pattern Recognition - Optimisation in Classification
- Fuzzy Clustering - Constrained Classification
- Nonhierarchical Methods

**Data Analysis**
- Textual Data Analysis - Categorical Data Analysis
- Multidimensional Scaling - Symbolic Data Analysis
- Robust Data Analysis - Correspondence Analysis
- Multivariate Data Analysis - Principal Components Analysis
- Time Series Analysis - Factor Analysis
- Regression Trees - Multiway Data Analysis
- Spatial Data Analysis

**Application of Classification or Data Analysis in:**
- Social Sciences - Industry
- Behavioural Sciences - Medicine
- Finance - Genome Analysis
- Management - Archaeology
- Marketing - Image Analysis
- Environmental Sciences - Risk Analysis
- Ecology - Cognition
- Biology - Quality Control

**CALL FOR PAPERS**

The researchers interested in classification, data analysis and related methods are encouraged to participate in the conference. Persons wishing to present a paper (as oral presentation or poster) should send:
- a title;
- an abstract for a book of abstracts of no more than 200 words;
- name(s) and affiliation(s) of the author(s);
- keywords.

... to the Chair of the Scientific Program Committee, Krzysztof Jajuga, before February 15, 2002. Please send this information to the following e-mail address: ifcs2002@credit.ae.wroc.pl

The abstract should follow the guidelines that can be found at the conference web site.

Please classify the paper in one of the categories summarised in the topics section of this Call for Papers. By default it will be assumed that the author wish to present orally. However, if there are too many oral presentations some papers may be deferred to the poster session. In case the author prefers a poster presentation, he/she should specifically indicate this wish under the abstract. If there is more than one author, please indicate who will present the talk. Each person is permitted to present only one contributed paper. This does not prevent an invited session speaker from presenting his/her contribution nor does it limit
the number of papers on which one may be listed as the co-author. The conference language is English.

**Proceedings**

The proceedings volume with a selection of papers will be published by Springer and it will be available at the conference. Authors wishing to submit their papers for publication in these proceedings should send a file containing LaTeX source of the paper to the Chair of the Scientific Program Committee, Krzysztof Jajuga, before November 1, 2001 at the e-mail address: ifcs2002@credit.ae.wroc.pl

The paper should not exceed six pages in length and should follow the guidelines that can be found at the conference web site. Publication in the proceedings volume is conditional to receiving the conference fee before February 15, 2002.

**Deadlines and Time Schedule**

November 1, 2001 - Deadline for the submission of manuscript for the proceedings
December 31, 2001 - Notification of the acceptance of the manuscript
January 31, 2002 - Deadline for sending in the revised manuscript
February 15, 2002 - Deadline for the submission of abstract
February 15, 2002 - Deadline for the conference fee payment for the authors of papers to be published in the proceedings
March 31, 2002 - Notification of the acceptance of the abstract

**Location: Cracow, Poland**

Cracow is the most famous and arguably the most beautiful city in Poland. It is among those rare cities of Europe where the medieval town is still the vital centre of the modern metropolis.

The cultural and social life is concentrated around the largest medieval market square in Europe. From the window of the tower of St. Mary Church a bugle call is played every hour. The church is famous for its masterpiece of late gothic art - the altar of St. Mary. Until the 17th century Cracow was the capital of Poland. Most of the Polish kings are buried here in the cathedral, next to the magnificent Wawel Castle. Cracow is the city of science and education with over 100,000 students, and more than 600 years old Jagiellonian University. The nearby 700 years old Wieliczka Salt Mine is on the First World Culture and Natural Heritage List of UNESCO.

**Hotels**

A list of hotels can be found on the conference web site. Information and reservation is available at: First Class Travel Office, Zwierzyniecka 29, Poland, 31-105 Krakow, Tel/fax + 48 12 4311385 e-mail: fckrakow@firstclass.com.pl
Guidelines for authors - IFCS-2002

To download required LaTeX files (guidekr.tex, guidekr.dvi, guidekr.ps, prockrak.tex) use that link guidekr.ziphttp://ifcs2002.ae.krakow.pl/prockrak.cls