FROM THE EDITOR

This issue contains seven articles on **sampling methods,** five **other articles**, three **reports,** an **obituary** and **acknowledgements** of referees of volume 6.

There are following seven articles devoted to some problems of sampling surveys:

1. ***Ranked Set Sampling and Its Applications in Finite Population Studies*** (by B. Kowalczyk from Poland). Several authors have studied various problems concerning ranked set sampling (RSS) design when sampling is implemented from an infinite population. In the present paper emphasis is placed on the implementation of RSS procedure from finite populations. The author examines basic properties of a population mean and total estimator under RSS using unified finite population theory. The author also conducts several simulation studies based on real data, which illustrate possibilities of applications of RSS design and point out a new aspect of RSS, which has not been analyzed so far. Ranking (semi-ranking) on a basis of grouped data is the new aspect of RSS presented in this paper. This approach seems to have many applications in finite population practice.

2. ***Multivariate Sample Allocation Problem in Two Schemes of Two-Stage Sampling*** (by M. Kozak from Poland). The paper contains considerations on a multivariate sample allocation in two-stage sampling. The author proposes the formulas of the sample allocation in two schemes of the multivariate two-stage sampling, i.e., *SI*, *SI*, and a special case of *pps*, *SI*, with the same sample sizes of secondary sampling units from primary sampling units. In both cases a self-weighting design is used.

3. ***A General Family of Estimators for Estimating Population Mean Using Two Auxiliary Variables in Two-Phase Sampling*** (by H.P. Singh, L. N. Upadhyaya and P. Chandra from India). A general family of estimators for estimating the population mean of the variable under study which make use of two auxiliary variables in two-phase sampling is proposed. Under simple random sampling without replacement (SRSWOR) scheme, the expressions of bias and mean-squared error (MSE) up to first order of approximation are derived. The family of estimators in its optimum case is discussed. It has been shown that the proposed family is more efficient than those considered by Srivastava (1970), Chand (1975), Srivastava et al (1990), Singh et al. (1994 a), Singh and Upadhyaya (1995), Singh (2001), Upadhyaya and Singh (2001) and the usual unbiased estimator and usual two-phase sampling estimator. An empirical study is carried out to illustrate the superiorperformance of the constructed estimator compared to others.

4. ***Estimation of Population Mean with Known Coefficient of Variation Under Optional Randomized Response Model Using Scrambled Response Technique*** (by H.P. Singh and N. Mathur from India). In this paper the authors have suggested an estimator of population mean when the coefficient of variation is available; and its properties are studied. An estimator for population mean is also suggested when a prior value of population mean is available along with its properties. Numerical illustrations are given to judge the merits of the suggested estimators over Gupta et. al. (2002) estimator.

5. ***An Optimum Size of Sampling Units for Tajikistan Household Survey*** (by Shyam Upadhyaya from Nepal). This article presents an attempt to find an optimum size of sampling units for the household survey in Tajikistan. The cost analysis showed that the number of PSU's could be significantly increased from the same budget if the necessary information on required parameters were known in advance.

6. ***A New Approach to πPs Sampling Scheme-I*** (by M. Saxena and A. Saxena from India), In the paper a sampling scheme using auxiliary variable for cluster sampling has been proposed. The auxiliary variable is used in selecting the elements in clusters. The efficiency of the proposed sampling scheme is compared to the usual cluster sampling with equal probabilities of selection as well as to SRSWOR. It has been observed that of all the schemes proposed the sampling scheme which makes use of πPS approach is the best.

7. ***Manipulation of the Irresolute Voters in Political Sample Surveys: A Mathematical Process*** **(**by N. Farmakis from Greece). In sample surveys designed to search for voting intentions, the problem of the manipulation of the irresolute voters (voters who have not decided about their voting, "floating" voters) arises. A method for managing their distribution is proposed in this paper. The method is very simple and is based on certain empirical coefficients from the Greek political experience of the last 25 years.

The second part of this issue under the title **Other Articles** contains five articles devoted to different topics:

8. ***Stochastic Orders and Classes of Prior Distributions*** (by M. Męczarski from Poland). The usual stochastic order and the dispersive order are considered in the set of all probability distributions as tools to construct classes of prior distributions. These classes are widely exploited in Bayesian robust statistical methods.

9. ***Towards Further Development of the Polish Consumer Price Statistics*** (by A. Głuchowska from Poland).The article summarises surveys on retail and consumer prices which used to be carried out by the official statistics in Poland and it describes also main aspect of the methodology of consumer price indices applied currently in the country. The author stressed that the biggest

step in this field was made in the second part of the nineties when the EU harmonised methodology was gradually implemented in Polish statistics. The article illustrates progress made in the Polish CPI and indicates those aspects of it which need further improvement. In concluding remarks the author emphasizes the differences concerning some other aspects: system of weights, coverage, product scope, aggregation and data publication, could be formulated so as to meet the needs of the main users of data.

10. ***The Information Content of Analysts' Recommendations: The Empirical Evidence from the Warsaw Stock Exchange*** (by H. Gurgul and P. Majdosz from Poland). In this article the authors investigate the impact of analysts' recommendations on stock prices using data from the Polish stock market. To improve market transparency, the authors' interest in this subject is limited to those companies listed on the Warsaw Stock Exchange (WSE) whose shares are the most liquid. Applying event study methodology, they define abnormal return as a difference between actual return of a given firm and expected return generated by the market model. The results show that *buy* as well as *hold* recommendations have no informational content for the investors operating on the WSE. Robustness of the authors' findings is confirmed by using bootstrap technique. This study is the first of this kind for the Polish stock market, and according to authors' opinion, also the first for the emerging markets of Central and Eastern Europe.

11. ***Care Needs and Clinical Characteristics of In-Patients in Southern Estonia*** (by G. Ljunggren from Sweden, K. Saks, E-M. Tiit from Estonia, B. Osberg from Sweden, and T. Kivastik from Estonia).The aim of this survey was to determine the structure of comprehensive needs of patients admitted to emergency wards. The sample consisted of all in-patients in all 15 emergency hospitals in the region of south-east Estonia.

12. ***The Factor Analysis in Research of the Spatial Differentiation of Poland's Economic Structure*** (by A. Malina from Poland). The paper presents the results of empirical studies in which the method of factor analysis was applied. The analysis has allowed the author to: a) separate the main factors on the basis of an initially given set of primary characteristics. The factors can be interpreted as determinants of development, b) rank voivodships in Poland to the identified main factors, c) determine groups of voivodships similar in the development level as well as in the economic structure.

There are also three **Reports:**

(i)*Celebrating Statistics — International Conference in Honour of Sir David Cox on the Occasion of His 80th Birthday*, Neuchâtel, 14—18 July 2004 (prepared by B.Kowalczyk);

(ii) *27th CIRET Conference*: Review of Selected Papers, Warsaw, Poland, 15—17 September 2004 (prepared by I. Zagoździńska);

(iii) *XXIIIrd Conference on Multivariate Statistical Analysis* (MSA 2004) & *Symposium on Statistics to Commemorate the 110th Anniversary of Jerzy Neyman's Birthday*, Łódź, Poland, 8—10 November 2004 (prepared by M. Pekasiewicz)

The editor announces with the deepest sorrow that Professor Aleksander Ryszard Wójcik, an Associate Editor of our journal, passed away in December 2003. **Obituary** of Professor A. Ryszard Wójcik is given in this issue.

The issue is concluded with the Acknowledgements of referees of Volume 6.

Jan Kordos

The Editor

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1031—1046

RANKED SET SAMPLING AND ITS APPLICATIONS IN FINITE POPULATION STUDIES

Barbara Kowalczyk[1]

ABSTRACT

Several authors have studied various problems concerned with ranked set sampling RSS design when sampling is implemented from an infinite population. In the present paper emphasis is put on the implementation of RSS procedure from finite populations. The author examines basic properties of a population mean and total estimator under RSS by using unified finite population theory. The author also conducts several simulation studies based on real data, which illustrate possibilities of applications of RSS design and point out a new aspect of RSS, which has not been analyzed so far. The new aspect of RSS presented in this paper is ranking (semi-ranking) on a basis of grouped data. This appronch seems to have many applications in finite population practice.

***Key words*:** order statistics, ranked set sampling, finite populations.

1. Introduction

The method of sampling based on ranked sets was first suggested by McIntyre (1952) as a way to estimate mean pasture and forage yields. Standard ranked set sampling RSS procedure involves **drawing from an infinite population** n random samples with n elements in each sample, which yields $n * n$ independent random variables Y_{11}, Y_{12}, ..., Y_{1n}, Y_{21}, Y_{22}, ..., Y_{2n}, ..., Y_{n1}, Y_{n2}, ..., Y_{nn} all having the same cumulative distribution function. The next step of RSS is to rank each sample (without actual measurements) with respect to the variable of interest Y:

$$\begin{array}{lccccc} 1 & Y_{1(1:n)} & Y_{1(2:n)} & Y_{1(3:n)} & \cdots & Y_{1(n:n)} \\ 2 & Y_{2(1:n)} & Y_{2(2:n)} & Y_{2(3:n)} & \cdots & Y_{2(n:n)} \\ 3 & Y_{3(1:n)} & Y_{3(2:n)} & Y_{3(3:n)} & \cdots & Y_{3(n:n)} \end{array}$$

[1] Warsaw School of Economics, bkowal@sgh.waw.pl

$$\begin{matrix} \vdots & & & & & \\ n & Y_{n(1:n)} & Y_{n(2:n)} & Y_{n(3:n)} & \dots & Y_{n(n:n)} \end{matrix}$$

From the first sample the element with the smallest rank is chosen for the actual measurement. From the second sample the element with the second smallest rank is chosen. The process is continued until the element with the largest rank from the *n*th sample is chosen. This procedure yields a total number of *n* elements chosen to be measured, one from each sample. The elements $Y_{1(1:n)}$, $Y_{2(2:n)}$, ..., $Y_{n(n:n)}$ constitute the ranked set sample. It has to be emphasized that $Y_{1(1:n)}$, $Y_{2(2:n)}$, ..., $Y_{n(n:n)}$ are independent as contrasted with the usual order statistics $Y_{1:n}$, $Y_{2:n}$, ..., $Y_{n:n}$ of a sample of *n* elements.

The idea of ranked set sampling is based on the following statement. There are many real life situations, in agricultural, environmental, biological, social and economic studies, where the variable of interest is not easily measurable, but it can be relatively easily ranked (order) at no additional cost or very little cost. In other words, in many situations there is enough information to do ranking without quantifying the units. The ranking can be done on the basis of visual inspection, prior information, earlier sampling episodes or other rough methods not requiring actual measurement. If there is a related variable which is readily observable and can be easily ranked, and which we feel is correlated with the variable of interest, the ranking can also be done on the basis of this auxiliary variable.

The standard example illustrating the matter under discussion is the following. If interest lies in estimating the mean height of trees, then measuring the height of the sampled trees could pose a problem, but it would be relatively easy to rank small sets of trees on the basis of visual inspection of their heights.

McIntyre (1952), who introduced ranked set sampling, did not supply any mathematical theory to support his suggestion. Takahasi and Wakimoto (1968) independently arrived at the same results and supplied the necessary statistical theory. They showed that mean of a ranked set sample $\overline{Y}_{RSS} = \frac{1}{n}\sum_{i=1}^{n} Y_{i(i:n)}$ is an unbiased estimator of the population mean and is more efficient than the usual sample mean $\overline{Y}$ under simple random sampling, that is when both estimators are constructed on the basis of the same number *n* of actual measurements, $\overline{Y}_{RSS}$ has smaller variance, which is given by a formula

$$Var(\overline{Y}_{RSS}) = \frac{1}{n}\left(\sigma_Y^2 - \frac{1}{n}\sum_{i=1}^{n} \left(E(Y_{i:n}) - \mu_Y \right)^2 \right).$$

McIntyre (1952) and Takahasi and Wakimoto (1968) assumed perfect ranking of the elements. Accurate ranking is the most difficult part of implementation of ranked set sampling procedure. Dell and Clutter (1972) considered the case in which the ranking might be done with errors, i.e. the

quantified element from the *i*th sample may not be the *i*th order statistic, but rather the *i*th "judgment order statistic". Let us denote the "judgment order statistic" and the "true order statistic" by $Y_{[i:n]}$ and $Y_{(i:n)}$ respectively. Dell and Clutter (1972) used the model

$$Y_{[i:n]} = Y_{(i:n)} + \varepsilon_i .$$

They assumed that $Y_{(i:n)}$ and ε_i are independent and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Dell and Clutter showed that the estimator

$$\bar{Y}_{RSS(e)} = \frac{1}{n}\sum_{i=1}^{n} X_{i[i:n]} ,$$

based on the given above model is an unbiased estimator of the population mean and is more efficient than the usual sample mean $\bar{Y}$ under simple random sampling, i.e.

$$Var(\bar{Y}_{RSS(e)}) \le Var(\bar{Y}) ,$$

with equality occurring only if the ranking is no better than random.

Stokes (1977) considered the situation when an auxiliary variable may be used to acquire a ranked set sample of the variable of interest. To carry out the ranking *n* samples are drawn independently, each having *n* units. From the first sample of size *n*, the *Y* associated with the smallest *X* is measured. From the second sample of size *n* the *Y* associated with the second smallest *X* is measured. We continue this way until the *Y* associated with the largest *X* from the *n*th sample is chosen for measurement. Stokes showed the following. Let the regression of *X* and *Y* is linear and the standardized variates $(X - \mu_X)/\sigma_X$ and $(Y - \mu_Y)/\sigma_Y$ have the same cumulative distribution function. Then the estimator $\bar{Y}_{RSS(a)}$ obtained on the basis of an auxiliary variable is more efficient then the one $\bar{Y}$ of standard simple random sampling and relative efficiency is given by

$$\frac{Var(\bar{Y}_{RSS(a)})}{Var(\bar{Y})} = \left(1 - \frac{\rho_{XY}^2}{n}\sum_{i=1}^{n}\left(\frac{E(X_{i:n}) - \mu_X}{\sigma_X}\right)^2\right).$$

Stokes (1980) also proposed an estimator of the population variance based on a ranked set sample and (1986) gave a useful review of some theoretical aspects of RSS. Mutlak and McDonald (1990a and 1990b) developed the RSS theory when the experimental units are selected with size-biased probability of selection. Yu and Lam (1997) proposed regression-type RSS estimator of the population mean.

The presented above method is the so-called one cycle RSS procedure. The problem of *m* cycles of RSS procedure is less important from the presented in this paper point of view and will not be discussed here. The interested reader is encouraged to go to Patil, Sinha and Taillie (1994). All the theory presented above has been almost exclusively confined to infinite populations. In the next section some unified finite population theory will be given.

2. Ranked Set Sampling from Finite Populations

When finite populations are taken into account, the need arise for the precise definition of a ranked set sampling procedure and examination of its properties by using finite population methods. In the present section the author describes ranked set sampling from finite populations, making distinction between with and without replacement designs.

Let us consider a finite population of N units $U = \{1,2,...,N\}$. The value of a characteristic Y for the jth population element is denoted by Y_j, j=1,2,...,N. The Y_j are treated as unknown but **constant** values (nonrandom).

From the given finite population of N elements a sample of n elements is drawn by simple random sampling without replacement SRSWOR. The drawing is repeated independently n times. So we have n independent samples (sets) of size n each. **It has to be emphasized that in each set separately distinct elements appear (within each set sampling is without replacement) but some elements that appear in one sample may also appear in some other sample (different samples are drawn independently from the entire population of *N* elements).** Next we rank each sample (set) without actual measurements. We choose for the final (actual) sample the element with the smallest rank from the first sample, the element with the second smallest rank from the second sample and so on until we choose the element with the largest rank from the nth sample. There is a total of n elements for actual measuring, one from each ordered sample.

Let y_{ij}, i=1,2,...,n, j=1,2,...,n denote the value of Y for the unit drawn in the ith sample (set) and in the jth draw. It is easily seen that y_{ij} is a random variable which can take values Y_1, Y_2, Y_3,..., Y_N, with probability $\frac{1}{N}$ each. We can present the drawings in the following diagram:

$$
\begin{array}{ll}
1 & y_{11}\ \ y_{12}\ \ y_{13}\ \cdots\ y_{1n} \\
2 & y_{21}\ \ y_{22}\ \ y_{23}\ \cdots\ y_{2n} \\
3 & y_{31}\ \ y_{32}\ \ y_{33}\ \cdots\ y_{3n} \\
\cdots & \\
n & y_{n1}\ \ y_{n2}\ \ y_{n3}\ \cdots\ y_{nn}
\end{array}
$$

Within each set sampling is without replacement (SRSWOR) and different sets are drawn independently. Let $y_{i(1:n)}, y_{i(2:n)}, \ldots, y_{i(n:n)}$ denote the corresponding order statistics of $y_{i1}, y_{i2}, \ldots, y_{in}$. After ranking we have:

$$\begin{array}{lccccc} 1 & \left(y_{1(1:n)}\right) & y_{1(2:n)} & y_{1(3:n)} & \cdots & y_{1(n:n)} \\ 2 & y_{2(1:n)} & \left(y_{2(2:n)}\right) & y_{2(3:n)} & \cdots & y_{2(n:n)} \\ 3 & y_{3(1:n)} & y_{3(2:n)} & \left(y_{3(3:n)}\right) & \cdots & y_{3(n:n)} \\ \cdots & & & & & \\ n & y_{n(1:n)} & y_{n(2:n)} & y_{n(3:n)} & \cdots & \left(y_{n(n:n)}\right) \end{array}$$

Elements $y_{1(1:n)}, y_{2(2:n)}, \ldots, y_{n(n:n)}$ constitute the ranked set sample. Quantified units are taken in parenthesis in the above diagram. It has to be emphasized that $y_{1(1:n)}, y_{2(2:n)}, \ldots, y_{n(n:n)}$ are independent as contrasted with the usual order statistics $y_{1:n}, y_{2:n}, \ldots, y_{n:n}$ of a sample of n elements selected by SRSWOR from a finite population which are correlated.

The mean of the ranked set sample is denoted by $\bar{y}_{RSS}$:

$$\bar{y}_{RSS} = \frac{1}{n}\sum_{i=1}^{n} y_{i(i:n)} \,. \tag{1}$$

Now we give properties of a ranked set sample mean $\bar{y}_{RSS}$ from a finite population, when ranked set sampling is defined as above.

1. The mean $\bar{y}_{RSS}$ of a ranked set sample from a finite population is an unbiased estimator of the population mean $\bar{Y}$.

   Proof:

$$E(\bar{y}_{RSS}) = \frac{1}{n}\sum_{i=1}^{n} E(y_{i(i:n)}) = \frac{1}{n}\sum_{i=1}^{n} E(y_{i:n}) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_{i:n}\right) = E(\bar{y}) = \bar{Y}$$

2. The variance of the mean $\bar{y}_{RSS}$ of a ranked set sample from a finite population is given by a formula:

$$D^2(\bar{y}_{RSS}) = \frac{1}{n}\left\{\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})]-\bar{Y}\right)^2\right\}, \tag{2}$$

   where

$$S^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(Y_j - \bar{Y}\right)^2 \,.$$

   Proof:

$$D^2(\bar{y}_{RSS}) = \frac{1}{n^2}\sum_{i=1}^{n} D^2(y_{i(i:n)}) = \frac{1}{n^2}\sum_{i=1}^{n}\left(E(y_{i(i:n)}^2) - [E(y_{i(i:n)})]^2\right) =$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} E(y_{i:n}^2) - \sum_{i=1}^{n}[E(y_{i:n})]^2\right) = \frac{1}{n^2}\left(E(\sum_{i=1}^{n} y_{i:n}^2) - \sum_{i=1}^{n}[E(y_{i:n})]^2\right)$$

$$= \frac{1}{n^2}\left\{E\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n}[E(y_{i:n})]^2\right)\right\} = \frac{1}{n^2}\left\{\sum_{i=1}^{n}\frac{1}{N}\sum_{j=1}^{N} Y_j^2 - \left(\sum_{i=1}^{n}[E(y_{i:n})]^2\right)\right\}$$

$$= \frac{1}{n}\left(\frac{1}{N}\sum_{j=1}^{N} Y_j^2 - \frac{1}{n}\sum_{i=1}^{n}[E(y_{i:n})]^2\right) = \frac{1}{n}\left(\frac{1}{N}(N-1)S^2 + \bar{Y}^2 - \frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})]^2\right)\right) =$$

$$= \frac{1}{n}\left(\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})] - \bar{Y}\right)^2\right).$$

3. The mean $\bar{y}_{RSS}$ of a ranked set sample from a finite population is more efficient than the common sample mean $\bar{y}_{SRS}$ under simple random sampling with replacement. The percentage gain in efficiency is given by:

$$\frac{D^2(\bar{y}_{SRS}) - D^2(\bar{y}_{RSS})}{D^2(\bar{y}_{SRS})} \cdot 100\% = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{[E(y_{i:n})] - \bar{Y}}{\sigma}\right)^2 \cdot 100\%, \qquad (3)$$

where

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\left(Y_j - \bar{Y}\right)^2.$$

Proof:

For simple random sampling with replacement we have: $D^2(\bar{y}_{SRS}) = \frac{1}{n}\sigma^2 = \frac{1}{n}\left(1-\frac{1}{N}\right)S^2$, while for ranked set sampling from a finite population we have (2). Because of $\frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})] - \bar{Y}\right)^2 \geq 0$ we get:

$$D^2(\bar{y}_{RSS}) \leq D^2(\bar{y}_{SRS}).$$

Directly from the obtained properties 1-3 analogous conclusions might be drawn for an estimator $y_{RSS} = N\bar{y}_{RSS}$ of the population total.

4. A statistic of the form $y_{RSS} = N\bar{y}_{RSS}$ is an unbiased estimator of the population total Y.

5. The variance of the estimator y_{RSS} of the population total is given by a formula:

$$D^2(y_{RSS}) = \frac{N^2}{n}\left\{\left(1-\frac{1}{N}\right)S^2 - \frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})]-\bar{Y}\right)^2\right\}. \qquad (4)$$

6. The estimator y_{RSS} of the population total based on ranked set sample from a finite population is more efficient than the usual estimator $y_{SRS} = N\bar{y}_{SRS}$ based on simple random sampling with replacement. The percentage gain in efficiency is given by:

$$\frac{D^2(y_{SRS}) - D^2(y_{RSS})}{D^2(y_{SRS})}\cdot 100\% = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{[E(y_{i:n})]-\bar{Y}}{\sigma}\right)^2 \cdot 100\% \qquad (5)$$

Remark 1

The mean of a ranked set sample from a finite population is more efficient than the mean of usual simple random sampling (with replacement). In finite population practice simple random sampling without replacement SRSWOR is the most important one, especially when sample size is large compared with the population size. By using formula (2) and known formulas for SRSWOR we can show that: $D^2(\bar{y}_{RSS}) \le D^2(\bar{y}_{SRSWOR})$ if and only if the following inequality holds: $\frac{1}{N-1}\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2 \le \frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}\left([E(y_{i:n})]-\bar{Y}\right)^2$. But this result seems to have no practical importance. When the sample size is relatively small compared with the population size, approximate equality holds: $D^2(\bar{y}_{SRS}) \approx D^2(\bar{y}_{SRSWOR})$ and RSS should also be superior to SRSWOR. But when the sample size is large compared with *N* the fact that ranked set sampling allows repetitions may decrease its efficiency.

Remark 2

In the case of finite populations it seems to be suitable to define ranked set sampling without replacement **RSSWOR** procedure, which assures to have all distinct elements in a selected sample. From the given finite population of *N* elements we draw a sample of *n* elements by simple random sampling without replacement SRSWOR. We rank the sample and choose for the final sample the element with the smallest rank. We remove the chosen element from the population and draw from the population of *N*-1 elements a sample of *n* elements by SRSWOR. We rank the sample and chose for the final sample the element with the second smallest rank. Then we remove from the population the chosen element. Next we draw a sample of *n* elements by SRSWOR from the population

of N-2 elements and so on until we have N-(n-1) elements left in the population from which we draw the last sample of n elements by SRSWOR, rank it and choose for the final sample the element with the largest rank. Finally we have n elements in the sample, each distinct. So this procedure yields a sample in which elements appear **without replacement**. The procedure may be implemented for $n \le \frac{1}{2}(N+1)$. To study the properties of RSSWOR simulation studies are proposed.

3. Simulation Studies Based on Real Data

Example 1

Let us consider a population of Polish towns of 10,000—100,000 inhabitants. In Poland there are 368 such towns. Two different variables are taken into account:

Y – number of pharmacies (study variable),

X – number of inhabitants (auxiliary variable).

Data about inhabitants in a town (auxiliary information) are taken from the Polish Central Statistical Office. Data about pharmacies are used here with the kind permission of IMS Poland.

The total value of Y is $Y = 3623$ pharmacies. Population mean and coefficient of variation are respectively $\bar{Y} = 9.8451$ and $CV = 0.7174$. Population correlation coefficient between the study Y and auxiliary variable X is $\rho = 0{,}9036$.

Now we conduct on the given population a simulation study. The goal is to estimate the population total Y, which real value is actually known, which allows us to compare different sampling designs:

SRSWOR – simple random sampling without replacement

RSSWOR – ranked set sampling without replacement.

In the case of ranked set sampling, ranking is done on the basis of auxiliary (concomitant) variable X – number of inhabitants in a town. From the first sample of size n, the Y associated with the smallest X is measured; from the second sample of size n the Y associated with the second smallest X is measured; we continue this way until the Y associated with the largest X from the nth sample is chosen for measurement. So we actually measure a sample of n elements.

To compare different methods of estimation samples of n = 30 elements are drawn from the given population by SRSWOR and RSSWOR. Sampling is repeated 1000 times.

An estimate of the population total obtained for the jth, $j = 1,2,...,1000$ sampling is denoted by t_j and the real population total by Y. Simulation study results are presented in a table given below. The following notation is used:

- mean of 1000 repetitions: $\frac{1}{1000}\sum_{i=1}^{1000} t_i$
- bias: $\left(\frac{1}{1000}\sum_{i=1}^{1000} t_i\right) - Y$
- max absolute bias in a single draw: $\max_{i=1,\ldots,1000}\left\{|t_i - Y|\right\}$
- average absolute bias: $\frac{1}{1000}\sum_{i=1}^{1000} |t_i - Y|$
- MSE: $\frac{1}{1000}\sum_{i=1}^{1000}(t_i - Y)^2$

Table 1. Simulation results for two estimators of the total number of pharmacies

	y_{SRSWOR}	y_{RSSWOR}
Mean of 1000 repetitions	3488.86	3709.00
Bias	-134.14	86.00
Bias in %	-3.70%	2.37%
Max of 1000 repetitions	4796.27	4612.27
Min of 1000 repetitions	2367.47	3042.13
Range	2428.80	1570.13
Max absolute bias in a single draw	1255.53	989.27
Max absolute bias in a single draw in %	34.65%	27.31%
Average absolute bias	342.73	187.63
Average absolute bias in %	9.46%	5.18%
MSE	180779.79	57005.37
Root mean square error	425.18	238.76
Root mean square error in %	11.74%	6.59%
Number of repetitions in which an estimate gave the more accurate assessments	319	681

Source: own calculation

Ranked set sampling without replacement procedure on the basis of an auxiliary variable leads to the increase of the efficiency of the estimation. Gain in precision compared with standard simple random sampling without replacement is respectively 155.10 (4.28%) as far as average absolute bias is concerned and 186.42 (5.15%) as far as root mean square error is concerned. Relative gain in efficiency is respectively: $\frac{342.73 - 187.63}{342.73} \cdot 100\% = 45.25\%$ in the case of average

absolute bias and $\frac{425.18-238.76}{425.18}\cdot 100\% = 43.85\%$ in the case of root mean square error. Under RSSWOR the range of the results is by $\frac{2428.80-1570.13}{2428.80}\cdot 100\% = 35.35\%$ lower than under SRSWOR. Number of repetitions in which estimate gave more accurate assessments is higher in the case of RSSWOR more than twice: $\frac{681}{319} = 2.14$. All figures here point RSSWOR to be superior to SRSWOR. The gain in efficiency is very high.

Example 2

Let us consider a population of N = 284 municipalities into which Sweden is divided. Two different variables are taken into account:

Y – 1985 population in thousands (study variable),

X – 1975 population in thousands (auxiliary variable).

Data are taken here from Särndal, Swensson and Wretman (1992). The total value of Y is $Y = 8339$ thousand. Population mean and coefficient of variation are respectively $\overline{Y} = 29.36$ and $CV = 1.7559$. Population correlation coefficient between the study Y and auxiliary variable X is $\rho = 0{,}9984$.

As in the previous example we conduct a simulation study aiming at estimation of the population total Y (total population in Sweden in 1985) and comparison of different sampling schemes: SRSWOR and RSSWOR. In the case of ranked set sampling, ranking is done on the basis of auxiliary variable X – 1975 population. In the case of both sampling schemes sample size is n = 20. To compare different methods of estimation sampling is repeated 1000 times.

In the presented example the gain in precision is also achieved by using RSSWOR, relatively $\frac{2025.25-1513.25}{2025.25}\cdot 100\% = 25.28\%$ as far as average absolute bias is concerned. Under RSSWOR the range of the results is by $\frac{14001.20-11374.20}{14001.20}\cdot 100\% = 18.76\%$ lower than under SRSWOR. Number of repetitions in which estimate gave the more accurate assessments is higher in the case of RSSWOR almost twice $\frac{660}{340} = 1.94$.

Table 2. Simulation results for two estimators of the total 1985 population in Sweden

	y_{SRSWOR}	y_{RSSWOR}
Mean of 1000 repetitions	7470.46	8421.47
Bias	-868.54	82.47
Bias in %	-10.42%	0.99%
Max of 1000 repetitions	17664.80	16983.20
Min of 1000 repetitions	3663.60	5609.00
Range	14001.20	11374.20
Max absolute bias in a single draw	9325.80	8644.20
Max absolute bias in a single draw in %	111.83%	103.66%
Average absolute bias	2025.25	1513.25
Average absolute bias in %	24.29%	18.15%
MSE	6297406.69	5211883.06
Root mean square error	2509.46	2282,96
Root mean square error in %	30.09%	27.38%
Number of repetitions in which estimator gave the more accurate assessments	340	660

Source: own calculation

Example 3

In finite population practice, especially in social, market and consumer research, statisticians encounter situations in which ranking can be implemented on the basis of grouped data. The standard example is a questionnaire. If we wish to use questionnaires for implementation of ranking procedure, we have to face two basic problems. First problem is an existence of non-sampling errors. And the second problem is that on the basis of questionnaires we can only obtain grouped data, because in practice in a questionnaire in most cases we can only ask "what brackets are you in?". **So on the basis of that kind of data (grouped) we can only rank some groups of units between themselves but we can not rank units within groups.** Such situation has not been analyzed so far and seems to be very important in finite population practice.

Now we conduct a simulation study by using real data, when ranking is implemented on the basis of grouped questionnaire data. All data are used here with the kind permission of IMS Poland.

A subpopulation of 1000 Polish pharmacies was examined and sales of the pharmacies were measured. Such a research is very expensive because it requires a connection with pharmacies' computers to have the exact data. This method of data collection is characterized by lack of measurement and response errors. The method of data collection is very precise and at the same time is very expensive. Majority of cost is spent on the permission to use pharmacies' data. The cost of a questionnaire, compared with exact measurement by using real computer data is

insignificant. Apart from that, market research agencies very often conduct different questionnaire studies for many other purposes and possess additional questionnaire data. We also possess questionnaire data about sales for 2329 pharmacies. For 181 pharmacies data about sales are available both from the exact measurement on the basis of real computer data and from a questionnaire.

To study exclusively the problem of the use of grouped data in ranking implementation, pharmacies which gave incorrect answers in a questionnaire are omitted in the analysis. As a result we have 93 pharmacies, for which data are available both from the exact measurement on the basis of computer data (quantitative data) and from a questionnaire (only grouped data). All questionnaire data in analyzed pharmacies are devoid of errors, the only problem is that they cannot all be ranked, as in standard ranked set sampling procedure.

Let us treat these 93 pharmacies as a population. Variable under study is Y – average monthly sales. The real population mean and coefficient of variation is respectively $\bar{Y} = 125666.08$ zł, $CV = 0.6572$. Questionnaire groups are the following:

<50 thousand;
50-100 thousand;
101-150 thousand;
151-200 thousand;
201-300 thousand;
301-500 thousand;
> 500 thousand.

Now we conduct on the given population of $N = 93$ pharmacies a simulation study. Samples of n = 10 elements are drawn by SRSWOR and semi-ranked set sampling without replacement S-RSSWOR, which is defined as below. Actual measurements are taken on the basis of real computer data and semi-ranking is done on the basis of questionnaire grouped data. **We use the expression semi-ranked set sampling, because we actually do not rank all units, only groups of units between themselves and within each group order is random.**

S-RSSWOR

First we draw from the population of N elements a sample of n elements by SRSWOR. We rank the sample by using grouped data. When two or more elements in the sample belong to the same group they are placed within the sample group randomly. So we rank only groups of elements between themselves and within each group order is random. As a result we have a sample ordered from 1 to n (semi-ranked order). Next we choose for the final sample the element with the smallest rank, i.e. an element in the lowest group. To avoid repetitions we remove from the population the chosen element. Next we draw from the population of (N-1) elements a sample of n elements. We rank the sample by using grouped data and choose for the final sample the element with the second smallest rank and then we remove it from the population. We continue the

procedure until we draw from the population of (N-(n-1)) elements a sample of n elements, rank it by using grouped data and choose for the final sample the element with the largest rank, i.e. an element that stands in the highest group in the sample. We conduct actual measurements on the final sample of n elements.

To analyze the methods of sample selection sampling is repeated 1000 times. Simulation results are presented in a table below.

Table 3. Simulation results for two estimators of the population mean

	$\bar{y}_{SRSWOR}$	$\bar{y}_{S-RSSWOR}$
Mean of 1000 repetitions	121132.52	129591.35
Bias	-4533.57	3925.27
Bias in %	-3.61%	3.12%
Max of 1000 repetitions	217178.27	183869.87
Min of 1000 repetitions	71149.98	94610.59
Range	146028.29	89259.29
Max absolute bias in a single draw	91512.19	58203.79
Max absolute bias in a single draw in %	72.82%	46.32%
Average absolute bias	18471.35	12679.28
Average absolute bias in %	14.70%	10.09%
MSE	800295188	342910117
Root mean square error	28289.49	18517.83
Root mean square error in %	22.51%	14.74%
Number of repetitions in which an estimate gave the more accurate assessments	352	648

Source: own calculation

Semi-ranked set sampling without replacement implemented on the basis of questionnaire grouped data leads to the increase of the efficiency of the mean estimation. Gain in efficiency compared with SRSWOR is respectively: 5792.07 (4.61%) as far as average absolute bias is concerned and 9771.66 (7.77%) as far as root mean square error is concerned.

Relative gain in efficiency is: $\frac{18471.35 - 12679.28}{18471.35} \cdot 100\% = 31.36\%$ and $\frac{28289.49 - 18517.83}{28289.49} \cdot 100\% = 34.54\%$. Under S-RSSWOR the range of the results is by $\frac{146028.29 - 89259.29}{146028.29} \cdot 100\% = 38.88\%$ lower than under SRSWOR. Number of repetitions in which estimate gave more accurate assessments is 1.84 times higher in the case of S-RSSWOR.

Simulation results on semi-ranked set sampling implemented on real data are promising. When elements of a given sample can not all be ranked without actual measurements but can be grouped and these groups can be ranked between themselves, the sampling scheme based on semi-ranked samples have promise of applications instead of standard simple random sampling. This problem has not been analyzed before although the possibility of possessing grouped data on which ranking can be implemented seems typical for many applications. Studies on this method of sample selection should yet continue.

Practitioners must be also interested in the results of a simulation study when both grouped data and errors in answers are taken into account. For 181 pharmacies only 93 (51.85%) gave the true answer in a questionnaire. Detailed data obtained from questionnaires are given in a table below.

Table 4. Real sales vs questionnaire answers

Questionnaire Answers (in thous.)	**Real Sales (in thous.)**						
	< 50	**50 -100**	**101 - 150**	**151 - 200**	**201 - 300**	**301 - 500**	**>500**
	percentage						
<50	**35.29**	41.18	23.53				
50 -100	5.56	**54.17**	29.16	2.77	4.17	4.17	
101 – 150		12.77	**51.06**	29.79	6.38		
151-200		11.11	11.11	**51.85**	25.93		
201-300		7.14	7.14		**57.14**	28.58	
301 – 500			50			**50**	
>500		50					**50**

Source: own calculation

Real population mean of 181 pharmacies is $\bar{Y} = 138385.65$ zł and coefficient of variation $CV = 0.6016$. Simulation study results, when ranking is done on the basis of grouped questionnaire data burdened with errors in answers is presented in a table below. In the case of both sampling schemes sample size is $n = 10$.

As it is seen from the result table despite the presence of errors in questionnaire answers S-RSSWOR proved to be more efficient for estimating population mean. Gain in efficiency compared with SRSWOR is respectively 4193.32 (3.03%) as far as average absolute bias is concerned and 4753.56 (3.44%) as far as root mean square error is concerned. Relative gain in efficiency is $\frac{21750.73 - 17557.41}{21750.73} 100\% = 19.28\%$ and $\frac{27117.15 - 22363.59}{27117.15} 100\% = 17.53\%$ respectively. Number of repetitions in which estimate gave more accurate assessments is 1.34 times higher in the case of S-RSSWOR. The gain in efficiency is naturally lower compared with the previous example where all questionnaire answers were correct. But this gain in efficiency is achieved

although questionnaire data are burden with errors with almost 50% and only grouped data are available. Simulation results on presented real data are promising.

Table 5. Simulation results for two estimators of the population mean

	$\bar{y}_{SRSWOR}$	$\bar{y}_{S-RSSWOR}$
Mean of 1000 repetitions	137048.50	139114.81
Bias	-1337.15	729.16
Bias in %	-0.97%	0.53%
Max of 1000 repetitions	246041.84	234519.56
Min of 1000 repetitions	78286.08	84798.33
Range	167755.76	149721.23
Max absolute bias in a single draw	107656.19	96133.91
Max absolute bias in a single draw in %	77.79%	69.47%
Average absolute bias	21750.73	17557.41
Average absolute bias in %	15.72%	12.69%
MSE	735339858	500130228
Root mean square error	27117.15	22363.59
Root mean square error in %	19.60%	16.16%
Number of repetitions in which an estimate gave the more accurate assessments	428	572

Source: own calculation

REFERENCES

DELL, T.R. and CLUTTER, J.L. (1972), Ranked Set Sampling Theory with Order Statistics Background, *Biometrics* 28, 545—555.

McINTYRE, G. A. (1952), A Method of Unbiased Selective Sampling, Using Ranked Sets, *Australian Journal of Agricultural Research* 3, 385—390.

MUTLAK, H. A. And McDonald, L.L. (1990a), Ranked Set Sampling with Size Biased probability of Selection, *Biometrics* 46, 435—445.

MUTLAK, H. A. And McDonald, L.L. (1990b), Ranked Set Sampling with Respect to a Concomitant Variables and with Size Biased Probability of Selection, *Commun. Statististics* — Theory Math. 19 (1), 205—219.

PATIL, G.P., SINHA A.K. and TAILLIE C. (1994), Ranked Set Sampling, In: G.P. Patil and C.R. Rao, eds., *Handbook of Statistics*, Vol. 12, pp. 167—200.

SÄRNDAL, C.E., SWENSSON, B. i WRETMAN, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, pp.652—659.

STOKES, S.L. (1977), Ranked Set Sampling with Concomitant Variables, *Comm. Statististics* Theory Methods A 6, 1207—1211.

STOKES, S.L. (1980), Estimation of Variance using Judgment Ordered Ranked Set Samples, *Biometrics* 36, 35—42.

STOKES, S.L. (1986), Ranked Set Sampling. In: S. Kotz et al., eds., *Encyclopedia of Statistical Sciences*, Wiley, New York, 585—588.

TAKAHASI, K. And WAKIMOTO, K. (1968), On Unbiased Estimates of the Population Mean Based on the Sample Stratified Means of Ordering, *Ann. Inst. Statist. Math*. 20, 1—31.

YU. P. L. H. and LAM, K. (1997), Regression Estimator in Ranked Set Sampling, *Biometrics*, 53, 1070—1080.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1047—1054

MULTIVARIATE SAMPLE ALLOCATION PROBLEM IN TWO SCHEMES OF TWO-STAGE SAMPLING

Marcin Kozak[1]

ABSTRACT

The paper contains considerations on a multivariate sample allocation in two-stage sampling. The author proposes the formulas of the sample allocation in two schemes of the multivariate two-stage sampling, i.e., *SI*, *SI*, and a special case of *pps*, *SI*, with the same sample sizes of secondary sampling units from primary sampling units. In both cases a self-weighting design is used.

Key words: multivariate surveys, sample allocation, two-stage sampling.

1. Introduction

An estimation and allocation problem in a univariate two-stage sampling is quite good described in many copies, see e.g. Cochran (1977), Särndal *et al*. (1992) or Bracha (1996). Some new advanced results can be seen in Knottnerus (2003). However, most surveys carried out in the Central Statistical Office of Poland (CSO) are multivariate, i.e., they aim in the estimation of many parameters. In such surveys the univariate sample allocation methods are not efficient. Till now there were just few attempts to the problem of the multivariate sample allocation in the two-stage sampling; see Kokan and Khan (1967), Waters and Chester (1987), or Wywiał (2003).

The aim of the paper is to present the considerations on this problem. The appropriate formulas are presented in two schemes of two-stage sampling; first one is *SI*, *SI* (using denotation of Särndal *et al*., 1992), i.e., with simple random sampling without replacement on both stages, and second one *pps*, *SI*, i.e., with *pps* sampling (with sampling probability proportional-to-size with replacement sampling) on the first stage, and simple random sampling without replacement on the second stage. We will consider a special case of the latter, i.e., with the same

[1] Department of Mathematical Statistics and Experimentation, Warsaw Agricultural University, Nowoursynowska 159, 02-776 Warsaw, Poland. e-mail: marcin.kozak@omega.sggw.waw.pl

sample sizes of secondary sampling units from primary sampling units. In both schemes a self-weighting design is used.

2. Two-Stage Sampling

Consider a population U consisting of N units, divided into M separate groups $U_1,...,U_M$, $M > 2$. Hence,

$$\begin{gathered} \bigcup_{h=1}^{M} U_h = U, \\ U_h \cap U_g = \varnothing \text{ for } g \neq h. \end{gathered} \tag{1}$$

These groups are called the primary sampling units (PSUs). Each PSU, say hth $(h = 1,...,M)$, consists of N_h sampling units, the so-called secondary sampling units (SSUs). Because of (1), we can write

$$\sum_{h=1}^{M} N_h = N. \tag{2}$$

In particular section we will write appropriate formulas for a variance of an estimator of total for the considered scheme of two-stage sampling. Its form influences the allocation formula. But first, let us present some considerations on the multivariate sample allocation in two-stage sampling found in the literature.

3. Sample Allocation

For the sake of simplicity, let us assume that the parameters of interest are the totals of k variables, i.e., the k-dimensional vector **Y**. In such a case the optimal allocation for the particular variable $Y_i, i = 1,...,k,$ might be not optimal or even inappropriate for the other variables. Therefore we see the need for multivariate sample allocation method under cost restrictions.

A dual problem was studied by Kokan and Khan (1967). The authors considered a different form of our question, because they minimized the survey cost with respect to k variance constraints. But in many surveys carried out in Central Statistical Office the problem is formulated as it was mentioned above – minimize the precision of estimation with respect to the fixed cost of the survey. Obviously, these both problems are different, so the results of Kokan and Khan (1967) are impractical in our question.

A traditional approach to the multivariate allocation is to solve k individual optimization problems, (i.e., to apply the univariate optimal allocation for each ith, i=1,…,k, variable), and then choose the final allocation from among the individual optima (Waters and Chester, 1987). Such approach was used by Kelley (1971), and Schweigert and Sibert (1983). Unfortunately, it is not an optimal

solution in the multivariate sense, for it does not consider all possible allocations, just the limited set. Moreover, often it is not obvious which individual optimum is the best for the whole group of the variables.

One of the possible solutions is to use the method of Kokan and Khan (1967) iteratively. The problem would resolve itself into a searching of such constraints for particular k variances, that would enable to get the optimal precision of estimation with respect to the fixed survey cost. Such procedure would be rather time-consuming; moreover, it would be hard to manipulate the constraints for the k variances (i.e., to find the best constraints.)

Waters and Chester (1987) proposed a graphical method to solve the problem formulated by Kokan and Khan (1967). The results of their method are susceptible on various interpretations, and their method does not improve the method of Kokan and Khan (1967). Therefore it should be treated rather as a demonstrative, but not analytical, method leading to the direct results.

Let us give a look at a problem of studying a precision of estimation. In the univariate case it can be understood as the variance or coefficient of variation of the estimator. But in the multivariate surveys there is no such direct and natural criterion (Holmberg, 2002). Wywiał (2003) proposed to use such formula of a generalized variance:

$$D^2(\mathbf{Y}) = \sum_{i=1}^{k} a_i D^2(Y_i), \tag{3}$$

where a_i, i=1,…,k, is the weight of the ith variable given by $a_i = S_{Y_i}^2 \left(\sum_{j=1}^{k} S_{Y_j}^2 \right)^{-1}$, $S_{Y_i}^2$ is the variance of the variable Y_i, and $D^2(Y_i)$ is the variance of the estimator of the total of Y_i; under constraints

$$E(\hat{K}) = K, \tag{4}$$

where $E(\hat{K})$ is the expected value of the survey cost, and K is its fixed value.

Holmberg (2002) studied a problem of a choice of a multivariate criterion in designing a sample survey. He compared three forms of the criterion, including the generalized formula (3). Moreover, he did not define the weight a_i, i=1,…,k, as Wywiał, but treated it as the weight informing about the importance of the ith estimator. In Holmberg's opinion the best multivariate criterion is a sum of relative efficiency losses, not the generalized formula (3). Therefore, using the results of Holmberg, the objective function in our problem can be presented as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{k} \frac{D^2(Y_i) - d_i^2}{d_i^2}, \tag{5}$$

where **x** is a vector of the allocation parameters, $d_i^2 = D_{opt}^2(Y_i), i = 1,...,k,$ is the variance of estimators of total Y_i from the optimal univariate allocation for the *i*th variable.

In both our problems the vector **x** is two-dimensional. We will seek its optimal values that minimize the function (5).

The criterion (5) will be used in finding the optimal multivariate sample allocation in two following sampling schemes of the two-stage sampling: *SI*, *SI* and *pps*, *SI*.

4. *SI, SI* Two-Stage Sampling

Let us consider the simplest case of the two-stage sampling, i.e., with the simple random sampling without replacement (*SI*) on both sampling stages, and the self-weighting design. The self-weighting design results from the fact that each SSU has the same sampling probability (Cochran, 1977; Särndal *et al*., 1992; and Bracha, 1996), i.e.,

$$\pi_{hj} = f_1 f_2 = const, h = 1,...,M, j = 1,...,N_h, \tag{6}$$

where π_{hj} is the probability of sampling the *j*th SSU from the *h*th PSU, $f_1 = m/M$ is the sampling fraction on the first stage of sampling, and $f_2 = n_h / N_h = const,\ h = 1,...,M,$ is the sampling fraction on the second stage of sampling.

$N_{(g)}$, the PSUs' sizes, and $n_{(g)}$, the sample sizes from the PSUs, are the random variables getting values with probabilities depending on the sampling scheme on the first stage of sampling (Bracha, 1996). Therefore, the sample size *n*,

$$n = \sum_{g=1}^{m} n_{(g)}, \tag{7}$$

is the sum of the random variables, so it is also the random variable (Bracha, 1996).

In such a case the estimator *y* of the total *Y* takes the form (Bracha, 1996)

$$y = \frac{1}{f_1 f_2} \sum_{g=1}^{m} \sum_{i=1}^{n_{(g)}} y_{gi}, \tag{8}$$

where $y_{gi}, g = 1,...,m, i = 1,...,n_{(g)}$ is the value of *Y* in the *i*th SSU from the *g*th PSU. The variance of the estimator (8) is given by

$$D^2(y) = \frac{M^2}{m}\left[(1-f_1)S_1^2 + \frac{1-f_2}{f_2}\overline{N}S_2^2\right] = \\ = M^2\left[\left(\frac{1}{m}-\frac{1}{M}\right)S_1^2 + \frac{\overline{N}(1-f_2)}{mf_2}S_2^2\right], \tag{9}$$

where

$$S_1^2 = \frac{1}{M-1}\sum_{h=1}^{M}\left(Y_h - \overline{\overline{Y}}\right)^2,\ \overline{Y} = \frac{Y}{M} = \frac{1}{M}\sum_{h=1}^{M}N_h\overline{\overline{Y}}_h,\ \overline{\overline{Y}} = \frac{1}{N}\sum_{h=1}^{M}\sum_{j=1}^{N_h}Y_{hj},$$

$$S_2^2 = \frac{1}{N}\sum_{h=1}^{M}N_h S_{2h}^2,\ \ S_{2h}^2 = \frac{1}{N_h-1}\sum_{j=1}^{N_h}\left(Y_{hj} - \overline{\overline{Y}}_h\right)^2,\ \overline{N} = \sum_{h=1}^{M}N_h.$$

The cost function takes the form of (Bracha, 1996)

$$\hat{K} = k_0 + mk_1 + k_2\sum_{g=1}^{m}n_{(g)}, \tag{10}$$

where k_0 is the fixed cost of the survey, k_1 is the cost per primary sampling unit, and k_2 is the cost of sampling one secondary unit from the PSU.

Note that because of the random $n_{(g)}$ in (10), $\hat{K}$ is also the random variable with an expected value given by

$$E(\hat{K}) = k_0 + m(k_1 + f_2\overline{N}k_2). \tag{11}$$

Let us introduce the values m_{opt} and f_{2opt} that are optimal in the multivariate sense, i.e., they minimize the criterion (5) with respects to the fixed expected survey cost *K*. We can obtain these values using the Lagrange method. For the sake of simplicity, in the Lagrange function we omit the term in the variance (9) that is unimportant from the point of view of the method. The cost restriction takes the form

$$E(\hat{K}) = K. \tag{12}$$

Hence, the Lagrange function is given by

$$L = \sum_{i=1}^{k}\frac{M^2}{d_i^2}\left(\frac{S_{1i}^2}{m} + \frac{\overline{N}(1-f_2)S_{2i}^2}{mf_2}\right) + \lambda\left[K - k_0 - m(k_1 - f_2\overline{N}k_2)\right], \tag{13}$$

where $S_{1i}^2 = \frac{1}{M-1}\sum_{h=1}^{M}\left(Y_{ih} - \overline{Y}_i\right)^2,\ S_{2i}^2 = \frac{1}{N}\sum_{h=1}^{M}N_h S_{2ih}^2,\ S_{2ih}^2 = \frac{1}{N_h-1}\sum_{j=1}^{N_h}\left(Y_{ihj} - \overline{\overline{Y}}_{ih}\right)^2.$

Provided that

$$\sum_{i=1}^{k}\left(S_{1i}^2-\overline{N}S_{2i}^2\right)d_i^{-2}>0, \tag{14}$$

and assuming that m and f_2 can take real values, the optimal m and f_2 are given by

$$f_{2opt}=\sqrt{\frac{k_1\sum_i S_{2i}^2 d_i^{-2}}{k_2\sum_i\left(S_{1i}^2-\overline{N}S_{2i}^2\right)d_i^{-2}}}, \tag{15}$$

$$m_{opt}=\frac{K-k_0}{k_1+k_2 f_{2opt}\overline{\overline{N}}}. \tag{16}$$

5. *pps, SI* Two-Stage Sampling

Let us consider the *pps*, *SI* scheme, i.e., with probability proportional-to-size with replacement sampling (*pps*) on the first stage, and simple random sampling without replacement sampling (*SI*) on the second stage. Moreover, consider a special case of this scheme, i.e., with the same sample sizes n_0 of the SSUs from the PSUs, and with the self-weighting design.

In such a case the estimator of total *Y* takes the form (Bracha, 1996)

$$y'=\frac{N}{mn_0}\sum_{g=1}^{m}\sum_{i=1}^{n_0}y_{gi}=N\overline{\overline{y}}', \tag{17}$$

where $\overline{\overline{y}}'=\left(mn_0\right)^{-1}\sum_{g=1}^{m}\sum_{i=1}^{n_0}y_{gi}$. A variance of the estimator (17) is given by

$$D^2\left(y'\right)=\frac{N}{m}\sum_{h=1}^{M}N_h\left(\overline{\overline{Y}}_h-\overline{\overline{Y}}\right)^2+\frac{N}{mn_0}\sum_{h=1}^{M}\left(N_h-n_0\right)S_{2h}^2, \tag{18}$$

where $\overline{\overline{Y}}_h=\frac{1}{N_h}\sum_{j=1}^{N_h}Y_{hj}$, $\overline{\overline{Y}}=\frac{1}{N}\sum_{h=1}^{M}N_h\overline{\overline{Y}}_h$. The cost restriction takes the form

$$K=k_0+m\left(k_1+n_0k_2\right). \tag{19}$$

Note that the cost (19) is fixed, in the contrary to (10).
The Lagrange function is given by

$$L=\frac{N}{m}\sum_{i=1}^{k}d_i^{-2}\left(a_i+b_i\right)+\lambda\left[K-k_0-m\left(k_1+n_0k_2\right)\right], \tag{20}$$

where $a_i = \sum_{h=1}^{M}\left[N_h\left(\overline{\overline{Y}}_{ih} - \overline{\overline{Y}}_i\right)^2 - S_{2ih}^2 \right]$, and $b_i = \sum_{h=1}^{M} S_{2ih}^2 N_h$.

Provided that

$$\sum_{i=1}^{k} d_i^{-2} \sum_{h=1}^{M}\left[N_h\left(\overline{\overline{Y}}_{ih} - \overline{\overline{Y}}_i\right)^2 - S_{2ih}^2 \right] > 0, \tag{21}$$

and assuming that n_0 and m can take real values, the optimal n_0 and m are given by

$$n_{0opt} = \sqrt{\frac{k_1 \sum_{i=1}^{k} b_i d_i^{-2}}{k_2 \sum_{i=1}^{k} a_i d_i^{-2}}} = \sqrt{\frac{k_1 \sum_i d_i^{-2} \sum_h S_{2ih}^2 N_h}{k_2 \sum_i d_i^{-2} \sum_h \left[N_h\left(\overline{\overline{Y}}_{ih} - \overline{\overline{Y}}_i\right) - S_{2ih}^2 \right]}}, \tag{22}$$

$$m_{opt} = \frac{K - k_0}{k_1 + k_2 n_{0opt}}. \tag{23}$$

6. Concluding remarks

In two previous sections we have proposed the formulas for the optimal multivariate sample allocation in two cases of the two-stage sampling. The sum of the relative efficiency was used as the multivariate criterion of the optimality and as the objective function. Unfortunately, we had to make some assumptions; note that when the restrictions (14) and (21) are not satisfied, the proposed formulas cannot be used. In such a case the only way is to use the mathematical programming.

Several additional aspects of the sample allocation have to be discussed. The proposed methods lead to the allocations given by the real values. Hence, they have to be rounded in an appropriate way, according to the given allocation. Certainly, it makes the solution not optimal. Moreover, when designing a survey, we base on some estimators of the parameters (the auxiliary information) used in the allocation formulas; they can originate from censuses, registers, previous or pilot surveys. It causes the allocation is optimal for the allocation variables, not for the survey ones. But, if the quality of the auxiliary information is good, i.e., the used estimators do not differ too much from the parameters, the allocation would be near optimal. On the other hand, if this quality is bad, the allocation would be far from optimal. Therefore, we should always take care of the good quality of the auxiliary information which we want to use in the sample allocation.

Note that both considered two-stage sampling schemes, especially the *SI*, *SI* scheme, are rather rare in a survey practice. Hence, should we treat the presented results as the theoretical ones only? No, they have their practical meaning. The most often two-stage sampling scheme is the one with the *πps* (probability

proportional-to-size without replacement) sampling on the first stage. But we do not know how to allocate a sample between sampling stages for the *πps*, *SI* two-stage sampling. Therefore, we can employ the results obtained for the *pps*, *SI* scheme. If the evaluated sample fraction on the first sampling stage is small, and there are no too large PSUs (which could be drawn to the sample many times, because of the big sampling probability), such procedure might be approximately optimal.

REFERENCES

BRACHA, Cz. (1996). *Theoretical Basis of Survey Sampling*. PWN, Warsaw, Poland (in Polish).

COCHRAN, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

HOLMBERG, A. (2002). A Multiparameter Perspective on the Choice of Sampling Design in Surveys. *Statistics in Transition*, 5, 969—994.

KELLEY, J.C. (1971). Multivariate Oceanographic Sampling. *Mathematical Geology*, 3, 45—50.

KOKAN, A. R., KHAN, S. (1967). Optimum Allocation in Multivariate Surveys: an Analytical Solution. *Journal of the Royal Statistical Society*, B, 29, 115—125.

KNOTTNERUS, P. (2003). *Sample Survey Theory. Some Pythagorean Perspectives*. Springer-Verlag, New York.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

SHWEIGERT, J.F., SIBERT, J.R. (1983). Optimizing Survey Design for Determining Age Structure of Fish Stocks: an Example from British Columbia Pacific Herring (*Clupea harengus pallasi*). *Canadian Journal of Fisheries and Aquatic Sciences*, 40, 588—597.

WATERS, J. R., CHESTER, A. J. (1987). Optimal Allocation in Multivariate, Two-Stage Sampling Designs. *The American Statistician*, 41, 46—50.

WYWIAŁ, J. (2003). *Some Contributions to Multivariate Methods in Survey Sampling*. Prace Naukowe Akademii Ekonomicznej w Katowicach, Poland.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1055—1077

A GENERAL FAMILY OF ESTIMATORS FOR ESTIMATING POPULATION MEAN USING TWO AUXILIARY VARIABLES IN TWO-PHASE SAMPLING

Housila P. Singh[1], **Lakshmi N. Upadhyaya**[2] **and P. Chandra**[3]

ABSTRACT

A general family of estimators for estimating the population mean of the variable under study which make use of two auxiliary variables in two-phase sampling is proposed. Under simple random sampling without replacement (SRSWOR) scheme, the expressions of bias and mean-squared error (MSE) up to first order of approximation are derived. The family of estimators in its optimum case is discussed. It has been shown that the proposed family is more efficient than those considered by Srivastava (1970), Chand (1975), Srivastava et al (1990), Singh et al. (1994 a), Singh and Upadhyaya (1995), Singh (2001), Upadhyaya and Singh (2001) and the usual unbiased estimator and usual two-phase sampling estimator. An empirical study is carried out to illustrate the performance of the constructed estimator over others.

Key words: Family of estimators, auxiliary variables, bias, mean-squared error.

1. Introduction

Consider a finite population U=$\{U_1, U_2, ..., U_N\}$. Let y and x be the study and auxiliary variable, taking values y_i and x_i respectively for the $i-$th unit U_i. The problem of estimating the population mean $\min.\mathrm{MSE}(t_2) = S_y^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)-\left(\frac{1}{n}-\frac{1}{n'}\right)\rho_{xy}^2\right]$ of y when the population mean $\overline{X}$ of

[1] Professor, School of Studies in Statistics, Vikram University, Ujjain-456010.India.
[2] Professor, Department of Applied Mathematics, Indian School of Mines, Dhanbad — 826004. India.
[3] Department of Biostatistics, All India Institute of Medical Sciences, New-Delhi-110029. India.

x is known has been dealt with in literature quite extensively. However, in certain practical situations when $\bar{X}$ is not known a priori, the technique of two-phase or double sampling is used. The two-phase sampling happens to be a powerful and cost effective (economical) procedure for finding the reliable estimate in first phase sample for the unknown parameters of the auxiliary variable x and hence has eminent role to play in survey sampling, for instance, see Hidiroglou and Sarndal (1995, 98).

Allowing SRSWOR design in each phase, the two-phase sampling scheme is as follows:

(i) The first phase sample $s_{n'}$ ($s_{n'} \subset U$) of a fixed size n' is drawn to measure only x in order to formulate a good estimate of a population mean $\bar{X}$,

(ii) Given $s_{n'}$, the second phase sample s_n ($s_n \subset s_{n'}$) of a fixed size n is drawn to measure y only.

Let $\bar{x} = (1/n)\sum_{i \in s_n} x_i$, $\bar{y} = (1/n)\sum_{i \in s_n} y_i$ and $\bar{x}' = (1/n')\sum_{i \in s_{n'}} x_i$.

The two-phase sampling ratio estimator of $\bar{Y}$ is defined by

$$t_1^{(R)} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right) \tag{1.1}$$

which is further generalized by Srivastava (1970) as

$$t_2 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_1} \tag{1.2}$$

where α_1 is determined so as to minimize the MSE of t_2. For more general formulation see Srivastava (1981) and Bedi (1985).

Sometimes even if $\bar{X}$ is not known, information on a cheaply ascertainable variable z, closely related to x but compared to x remotely related to y, is available on all units of the population. For instance, while estimating the total yield of wheat in a village, the yield and area under the crop are likely to be unknown, but the total area of each farm may be known from village records or may be obtained at a low cost. Then y, x and z are respectively yield, area under wheat and area under cultivation. This type of situation has been briefly discussed among others by, Chand (1975), Kiregyera (1980, 84) and Srivenkataramana and Tracy (1989).

Replacing simply $\bar{x}'$, in $t_1^{(R)}$, by the ratio estimator $t_1^{(R)'} = \bar{x}' \frac{\bar{Z}}{\bar{z}'}$ Chand (1975) obtained a chain ratio-type estimator

$$t_3 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\bar{Z}}{\bar{z}'}\right) \tag{1.3}$$

where $\bar{Z}$ is the known population mean of z and $\bar{z}' = (1/n') \sum_{i \in s_{n'}} z_i$.

Later, several modifications of Chand's chain ratio-type estimator t_3 have been presented by various authors including Kiregyera (1980, 84), Srivastava et al. (1988, 89, 90), Sahoo and Sahoo (1993, 94, 99a, 99b, 2001), Sahoo et al. (1993, 94), Singh et al. (1992, 94a, 94b), Singh and Singh (1991, 97—98), Prasad et al. (1996), Singh and Upadhyaya (1995), Mukherjee et al. (1987), Upadhyaya and Singh (2001), Singh and Espejo (2000), Tripathi and Ahmed (1995), Singh and Gangele (1989—95), Tracy and Singh (1999), Ahmed (1998), Upadhyaya et al. (1990, 92), Singh (1993), Ahmed at al. (1994) and Singh (2001). Some chain ratio-type estimators for estimating finite population variance have been proposed by Gupta et al. (1992—93).

In this paper, under SRSWOR, we have suggested a general family of estimators for estimating the population mean $\bar{Y}$. The expressions of bias and MSE, up to the first order of approximation, have been obtained, which will enable us to obtain the said expressions for any member of this family. The general family, in its optimum case and up to the first order of approximation, has been discussed. Further some well known estimators have been demonstrated as particular members of this family.

2. The suggested family of estimators

We define a family of estimators of $\bar{Y}$ as

$$\mathrm{t} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_1}\left(\frac{a\bar{Z}+b}{a\bar{z}'+b}\right)^{\alpha_2}\left(\frac{a\bar{Z}+b}{a\bar{z}+b}\right)^{\alpha_3} \tag{2.1}$$

where $\alpha_i's$ $(i = 1,2,3)$ are unknown constants to be suitably determined and $a(\neq 0), b$ are either real numbers or the functions of the known parameters of the

second auxiliary variable z such as standard deviation (σ_z), coefficients of variation (C_z), skewness $(\beta_1(z))$ and kurtosis $(\beta_2(z))$; and $\bar{z} = (1/n)\sum_{i \in s_n} z_i$.

To obtain the bias and MSE of t, we write

$$\bar{y} = \bar{Y}(1+e_0),\ \bar{x} = \bar{X}(1+e_1),\ \bar{x}' = \bar{X}(1+e_1'),\ \bar{z} = \bar{Z}(1+e_2),\ \bar{z}' = \bar{Z}(1+e_2')$$

such that

$$\mathrm{E}(e_0) = \mathrm{E}(e_1) = \mathrm{E}(e_1') = \mathrm{E}(e_2) = \mathrm{E}(e_2') = 0$$

and

$$\mathrm{E}(e_0^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_y^2, \qquad \mathrm{E}(e_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2,$$

$$\mathrm{E}(e_1'^2) = \left(\frac{1}{n'} - \frac{1}{N}\right) C_x^2, \qquad \mathrm{E}(e_2^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_z^2,$$

$$\mathrm{E}(e_2'^2) = \left(\frac{1}{n'} - \frac{1}{N}\right) C_z^2, \qquad \mathrm{E}(e_0 e_1) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho_{xy} C_x C_y,$$

$$\mathrm{E}(e_0 e_1') = \left(\frac{1}{n'} - \frac{1}{N}\right) \rho_{xy} C_x C_y, \qquad \mathrm{E}(e_0 e_2) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho_{yz} C_y C_z,$$

$$\mathrm{E}(e_0 e_2') = \left(\frac{1}{n'} - \frac{1}{N}\right) \rho_{yz} C_y C_z, \qquad \mathrm{E}(e_1 e_1') = \left(\frac{\beta_1(z)\bar{Z} + \sigma_z}{\beta_1(z)\bar{z}' + \sigma_z}\right)^{\alpha_2},$$

$$\mathrm{E}(e_1 e_2) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho_{xz} C_x C_z, \qquad \mathrm{E}(e_1 e_2') = \left(\frac{1}{n'} - \frac{1}{N}\right) \rho_{xz} C_x C_z,$$

$$\mathrm{E}(e_1' e_2) = \left(\frac{1}{n'} - \frac{1}{N}\right) \rho_{xz} C_x C_z, \qquad \mathrm{E}(e_1' e_2') = \left(\frac{1}{n'} - \frac{1}{N}\right) \rho_{xz} C_x C_z,$$

$$\mathrm{E}(e_2 e_2') = \left(\frac{1}{n'} - \frac{1}{N}\right) C_z^2,$$

where $C_y^2 = S_y^2 / \bar{Y}^2$, $C_x^2 = S_x^2 / \bar{X}^2$, $C_z^2 = S_z^2 / \bar{Z}^2$,

$\rho_{xy} = S_{xy} / (S_x S_y)$, $\rho_{xz} = S_{xz} / (S_x S_z)$, $\rho_{yz} = S_{yz} / (S_y S_z)$,

$S_y^2 = \sum_{i \in U} (y_i - \bar{Y})^2 / (N-1)$, $S_x^2 = \sum_{i \in U} (x_i - \bar{X})^2 / (N-1)$, $S_z^2 =$

$\sum_{i \in U} (z_i - \bar{Z})^2 / (N-1)$,

$S_{xy} = \sum_{i\in U}(x_i - \overline{X})(y_i - \overline{Y})/(N-1)$, $S_{xz} = \sum_{i\in U}(x_i - \overline{X})(z_i - \overline{Z})/(N-1)$,

and $S_{yz} = \sum_{i\in U}(y_i - \overline{Y})(z_i - \overline{Z})/(N-1)$.

Expressing t in terms of $e's$, we have

$$t = \overline{Y}\,(1+e_0)(1+e_1)^{-\alpha_1}(1+e_1')^{\alpha_1}(1+\phi\, e_2')^{-\alpha_2}(1+\phi\, e_2)^{-\alpha_3} \tag{2.2}$$

where

$$\phi = \frac{a\overline{Z}}{a\overline{Z}+b}. \tag{2.3}$$

We assume that $|e_1|<1$, $|e_1'|<1$, $|\phi\, e_2'|<1$, and $|\phi\, e_2|<1$,

so that $(1+e_1)^{-\alpha_1}$, $(1+e_1')^{\alpha_1}$, $(1+\phi\, e_2')^{-\alpha_2}$ and $(1+\phi\, e_2)^{-\alpha_3}$ are expandable.

Expanding the right hand side of (2.2) and retaining terms up to second powers of $e's$, we have

$$\begin{aligned} t = \overline{Y}\Big[&1+e_0+\alpha_1(e_1'-e_1)-\phi(\alpha_2 e_2'+\alpha_3 e_2)+\alpha_1(e_0e_1'-e_0e_1)-\phi(\alpha_2 e_0 e_2'+\alpha_3 e_0 e_2) \\ &+\alpha_1\alpha_2\phi(e_1e_2'-e_1'e_2')+\alpha_1\alpha_3\phi(e_1e_2-e_1'e_2)+\alpha_2\alpha_3\phi^2 e_2e_2' \\ &+\alpha_1\left\{\frac{(\alpha_1-1)}{2}e_1'^2+\frac{(\alpha_1+1)}{2}e_1^2-\alpha_1 e_1e_1'\right\}+\frac{\phi^2}{2}\left\{\alpha_2(\alpha_2+1)e_2'^2+\alpha_3(\alpha_3+1)e_2^2\right\}\Big] \end{aligned} \tag{2.4}$$

Taking expectation of both sides in (2.4) and then subtracting $\overline{Y}$ from both sides, we get the bias of the estimator t, up to the first order of approximation, as

$$\begin{aligned} B(t) = \left(\frac{\overline{Y}}{2}\right)\Bigg[&\left(\frac{1}{n}-\frac{1}{n'}\right)\alpha_1 C_x^2(\alpha_1-2K_{yx}+1) \\ &+\left(\frac{1}{n'}-\frac{1}{N}\right)C_z^2\left\{\alpha_2\phi(\alpha_2\phi-2K_{yz}+\phi)-\alpha_3\phi(K_{xz}\alpha_1-\alpha_2\phi)\right\} \\ &+\left(\frac{1}{n}-\frac{1}{N}\right)\alpha_3\phi\, C_z^2\left\{\phi(\alpha_3+1)+2\alpha_1K_{xz}-2K_{yz}\right\}\Bigg] \end{aligned} \tag{2.5}$$

where $K_{yx} = \rho_{yx}C_y/C_x$, $K_{xz} = \rho_{xz}C_x/C_z$ and $K_{yz} = \rho_{yz}C_y/C_z$.

From (2.4) we have

$$(t - \overline{Y}) \cong \overline{Y}\left[e_0+\alpha_1(e_1'-e_1)-\phi\,(\alpha_2 e_2'+\alpha_3 e_2)\right] \tag{2.6}$$

Squaring both sides of (2.6) and then taking expectation, we get the MSE of the estimator t, up to the first order of approximation, as

$$\text{MSE(t)}=\overline{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)\left\{C_y^2+\phi\,\alpha_3 C_z^2(\alpha_3\,\phi+2\alpha_1 K_{xz}-2K_{yz})\right\}\right.$$
$$+\left(\frac{1}{n'}-\frac{1}{N}\right)C_z^2\,\phi\left\{\alpha_2(\alpha_2\phi-2K_{yz})+2\alpha_3(\alpha_2\phi-\alpha_1 K_{xz})\right\}$$
$$\left.+\left(\frac{1}{n}-\frac{1}{n'}\right)\alpha_1 C_x^2(\alpha_1-2K_{yx})\right] \qquad (2.7)$$

Minimization of (2.7) with respect to α_1, α_2 and α_3 yields their optimum values as

$$\left.\begin{aligned}\alpha_1&=\frac{(K_{yx}-K_{zx}K_{yz})}{(1-K_{xz}K_{zx})}=\alpha_{10}\ \text{(say)}\\ \alpha_2&=\frac{K_{xz}(K_{yx}-K_{zx}K_{yz})}{\phi\,(1-K_{xz}K_{zx})}=\alpha_{20}\ \text{(say)}\\ \alpha_3&=\frac{(K_{yz}-K_{yx}K_{xz})}{\phi\,(1-K_{xz}K_{zx})}=\alpha_{30}\ \text{(say)}\end{aligned}\right\} \qquad (2.8)$$

where $K_{zx}=\rho_{xz}C_z/C_x$.

Substitution of (2.8) in (2.7) yields the minimum value of MSE(t) as

$$\text{min. MSE(t)} = S_y^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)-\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2-\left(\frac{1}{n}-\frac{1}{n'}\right)R_{y.xz}^2\right] \qquad (2.9)$$

where $R_{y.xz}^2 = (\rho_{xy}^2+\rho_{yz}^2-2\rho_{xy}\rho_{yz}\rho_{xz})/(1-\rho_{xz}^2)$ is the multiple correlation coefficient of y on x and z.

Thus we have the following theorem.

Theorem 2.1. Up to terms of order n^{-1},

$$\text{MSE(t)} \geq S_y^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)-\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2-\left(\frac{1}{n}-\frac{1}{n'}\right)R_{y.xz}^2\right],$$

with equality sign holding good if

$$\alpha_1=\frac{(K_{yx}-K_{zx}K_{yz})}{(1-K_{xz}K_{zx})},\quad \alpha_2=\frac{K_{xz}(K_{yx}-K_{zx}K_{yz})}{\phi\,(1-K_{xz}K_{zx})},\ \text{and}\ \alpha_3=\frac{(K_{yz}-K_{yx}K_{xz})}{\phi\,(1-K_{xz}K_{zx})}.$$

Remark 2.1. It is to be mentioned that the proposed family of estimators t will attain the minimum MSE in (2.9) only when the exact optimum values of α_{i0}'s $(i=1,2,3)$ at (2.8) of α_i's $(i=1,2,3)$ are known. The optimum values of α_{i0}'s $(i=1,2,3)$ are functions of unknown population parameters such as K_{yx}, K_{xz}, K_{zx}, K_{yz} and the known constant ϕ. The values of K_{yx}, K_{xz}, K_{zx} and K_{yz} can be guessed quite accurately from the past data or experience gathered in due course of time, for instance, see Srivastava (1967), Reddy (1973), Prasad (1989, p. 391) and Murthy (1967, pp. 96-99). Also the prior values of K_{yx}, K_{xz}, K_{zx} and K_{yz} may be either obtained on the basis of the information from a most recent survey or by conducting a pilot survey, see, Lui (1990, p.3805).

3. Some members of the proposed family of the estimators t

The following scheme presents some of the important known estimators of the population mean which can be obtained by suitable choice of constants α_i's $(i=1,2,3)$, a and b.

Estimator	Values of				
	α_1	α_2	α_3	*a*	*b*
1. $t_0 = \bar{y}$, the mean per unit estimator .	0	0	0	-	-
2. $t_1^{(R)} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)$, the usual two-phase sampling ratio estimator.	1	0	0	-	-
3. $t_1^{(P)} = \bar{y}\left(\frac{\bar{x}}{\bar{x}'}\right)$, the usual two-phase sampling product estimator.	-1	0	0	-	-
4. $t_2 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_1}$, Srivastava (1971) estimator .	α_1	0	0	-	-

Estimator	Values of				
	α_1	α_2	α_3	a	b
5. $t_3 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\bar{Z}}{\bar{z}'}\right)$, Chand (1975) chain ratio-type estimator.	1	1	0	1	0
6. $t_4 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_1}\left(\frac{\bar{Z}}{\bar{z}'}\right)^{\alpha_2}$, Srivastava et al. (1990) estimator.	α_1	α_2	0	1	0
7. $t_5 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\bar{Z}+C_z}{\bar{z}'+C_z}\right)^{\alpha_2}$, Singh and Upadhyaya (1995) estimator.	1	α_2	0	1	C_z known
8. $t_6 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\beta_2(z)\bar{Z}+C_z}{\beta_2(z)\bar{z}'+C_z}\right)$, Upadhyaya and Singh (2001) estimator.	1	1	0	$\beta_2(z)$ known	C_z known
9. $t_7 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{C_z\bar{Z}+\beta_2(z)}{C_z\bar{z}'+\beta_2(z)}\right)$, Upadhyaya and Singh (2001) estimator.	1	1	0	C_z known	$\beta_2(z)$ known
10. $t_8 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\beta_2(z)\bar{z}'+C_z}{\beta_2(z)\bar{Z}+C_z}\right)$, Upadhyaya and Singh (2001) estimator.	1	-1	0	$\beta_2(z)$ known	C_z known
11. $t_9 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{C_z\bar{z}'+\beta_2(z)}{C_z\bar{Z}+\beta_2(z)}\right)$, Upadhyaya and Singh (2001) estimator.	1	-1	0	C_z known	$\beta_2(z)$ known

Estimator	Values of				
	α_1	α_2	α_3	*a*	*b*
12. $t_{10} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\beta_2(z)\bar{Z}+C_z}{\beta_2(z)\bar{z}'+C_z}\right)^{\alpha_2}$	1	α_2	0	$\beta_2(z)$ known	C_z known
13. $t_{11} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{C_z\bar{z}'+\beta_2(z)}{C_z\bar{Z}+\beta_2(z)}\right)^{\alpha_2}$	1	α_2	0	C_z known	$\beta_2(z)$ known
14. $t_{12} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\bar{Z}+\sigma_z}{\bar{z}'+\sigma_z}\right)^{\alpha_2}$, Singh (2001) estimator.	1	α_2	0	1	σ_z known
15. $t_{13} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\beta_1(z)\bar{Z}+\sigma_z}{\beta_1(z)\bar{z}'+\sigma_z}\right)^{\alpha_2}$, Singh (2001) estimator.	1	α_2	0	$\beta_1(z)$ known	σ_z known
16. $t_{14} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\beta_2(z)\bar{Z}+\sigma_z}{\beta_2(z)\bar{z}'+\sigma_z}\right)^{\alpha_2}$, Singh (2001) estimator.	1	α_2	0	$\beta_2(z)$ known	σ_z known

In addition to these estimators a large number of estimators can also be generated from the proposed family of estimators t at (2.1) just by putting suitable values of $\alpha_1, \alpha_2, \alpha_3$, a and b.

It is observed that the expressions of the first order approximation of bias and MSE/Variance of the given member of the family can be obtained by mere substituting the values of $\alpha_1, \alpha_2, \alpha_3$, a and b in (2.5) and (2.7) respectively.

A class of chain ratio-type estimators for $\bar{Y}$ is defined by

$$t_{15} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left[\frac{\psi(\phi_1(d))}{\psi(\phi_2(d))}\right] \tag{3.1}$$

which is due to Singh et al. (1994 a), where

$$\psi\{\phi_i(d)\} = \phi_i(d) + \{1-\phi_i(d)\}\frac{\overline{Z}}{\overline{z}'}, \qquad (i = 1,2)$$

$$\phi_1(d) = f\,B/(A + f\,B + C),$$

$$\phi_2(d) = C/(A + f\,B + C), \qquad f = n/N,$$

$$A = (d-1)(d-2), \qquad B = (d-1)(d-4),$$

$$C = (d-2)(d-3)(d-4)$$

and $d(>0)$ is an unknown constant.

However, motivated by Srivastava (1967) and Naik and Gupta (1991), we suggest a generalized version of Singh et al.'s (1994a) estimator t_{15} as

$$t_{16} = \overline{y}\left(\frac{\overline{x}'}{\overline{x}}\right)^{\alpha_1}\left[\frac{\psi\{\phi_1(d)\}}{\psi\{\phi_2(d)\}}\right]^{\alpha_2} \tag{3.2}$$

For $\alpha_1 = \alpha_2 = 1$, t_{16} reduces to Singh et al.'s estimator (1994 a) and more estimators can be generated from t_{16} just by substituting the suitable values of α_1, α_2 and d.

Up to the first order of approximation, the bias and MSE of t_{16} are respectively given by

$$\mathrm{B}(t_{16}) = \left(\frac{\overline{Y}}{2}\right)\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\alpha_2\phi^* C_z^2\left\{\phi^*\alpha_2 + 2K_{yz} - (\phi_1+\phi_2)\right\}\right.$$

$$\left. + \left(\frac{1}{n}+\frac{1}{n'}\right)\alpha_1(\alpha_1 - 2K_{yx} + 1)\,C_x^2\right], \tag{3.3}$$

$$\mathrm{MSE}(t_{16}) = \overline{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2 + \left(\frac{1}{n'}-\frac{1}{N}\right)C_z^2\phi^*\alpha_2(\alpha_2\phi^* + 2K_{yz})\right.$$

$$\left. + \left(\frac{1}{n}-\frac{1}{n'}\right)\alpha_1 C_x^2(\alpha_1 - 2K_{yx})\right] \tag{3.4}$$

where $\phi^* = (\phi_1 - \phi_2)$.

It is to be noted that the bias and MSE of the estimator belonging to the class t_{16} can be obtained from (3.3) and (3.4) respectively just by putting the suitable values of α_1, α_2 and d.

The MSE (t_{16}) at (3.4) is minimized for

$$\left.\begin{aligned} \alpha_1 &= K_{yx} \\ \alpha_2 &= -\frac{K_{yz}}{\phi^*} \end{aligned}\right\} \tag{3.5}$$

and thus the resulting (minimum) MSE of t_{16} is given by

$$\text{min. MSE}(t_{16}) = S_y^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)-\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\rho_{yx}^2\right] \tag{3.6}$$

which equals to the minimum MSE of Srivastava et al. (1990) estimator t_4, i.e.,

$$\text{min. MSE}(t_{16}) = \text{min. MSE}(t_4) \tag{3.7}$$

Further, it can be shown that, to up the first order of approximation, the common minimum MSE of Singh et al. (1994) estimator t_{15}, Singh and Upadhyaya (1995) estimator t_5, the estimators t_{10}, t_{11}; and Singh (2001) estimators (t_{12}, t_{13}, t_{14}) is given by

$$\text{min.MSE}(t_j) = \bar{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2+\left(\frac{1}{n}-\frac{1}{n'}\right)C_x^2(1-2K_{yx})-\left(\frac{1}{n'}-\frac{1}{N}\right)K_{yz}^2C_Z^2\right] \tag{3.8}$$

(j = 5, 10, 11, 12, 13, 14, 15) .

The MSE/variance of the estimators $t_0 = \bar{y}$, $t_1^{(R)}$, t_3, up to the first order of approximation, are respectively given by

$$V(\bar{y}) = \left(\frac{1}{n}-\frac{1}{N}\right)\bar{Y}^2C_y^2 = \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2 \tag{3.9}$$

$$\text{MSE}\left(t_1^{(R)}\right) = \bar{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2+\left(\frac{1}{n}-\frac{1}{n'}\right)C_x^2(1-2K_{yx})\right] \tag{3.10}$$

$$\text{MSE}(t_3) = \bar{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2+\left(\frac{1}{n'}-\frac{1}{N}\right)C_z^2(1-2K_{yz})+\left(\frac{1}{n}-\frac{1}{n'}\right)C_x^2(1-2K_{yx})\right] \tag{3.11}$$

The minimum MSE of Srivastava (1971) estimator t_2, up to the first order of approximation, is given by

$$\min.\mathrm{MSE}(t_2) = S_y^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)-\left(\frac{1}{n}-\frac{1}{n'}\right)\rho_{xy}^2\right] \tag{3.12}$$

From (2.9), (3.6), (3.7), (3.8), (3.9), (3.10) and (3.11) we have

$$V(\bar{y}) - \min.\mathrm{MSE}(t) = S_y^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2+\left(\frac{1}{n}-\frac{1}{n'}\right)R_{y.xz}^2\right] > 0 \tag{3.13}$$

$$\mathrm{MSE}(t_1^{(\mathrm{R})}) - \min.\mathrm{MSE}(t) = S_y^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2+\left(\frac{1}{n}-\frac{1}{n'}\right)D\right] > 0 \tag{3.14}$$

$$\mathrm{MSE}(t_3) - \min.\mathrm{MSE}(t) = S_y^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\left(\frac{C_z}{C_y}-\rho_{yz}\right)^2+\left(\frac{1}{n}-\frac{1}{n'}\right)D\right] > 0 \tag{3.15}$$

$$\min.\mathrm{MSE}(t_j) - \min.\mathrm{MSE}(t) = S_y^2\left(\frac{1}{n}-\frac{1}{n'}\right)D > 0 \qquad (j = 5, 10 \text{ to } 15) \tag{3.16}$$

$$\min.\mathrm{MSE}(t_4 \text{ or } t_{16}) - \min.\mathrm{MSE}(t) = S_y^2\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{(\rho_{yz}-\rho_{xy}\rho_{xz})^2}{(1-\rho_{xz}^2)} > 0 \tag{3.17}$$

$$\min.\mathrm{MSE}(t_2) - \min.\mathrm{MSE}(t) = S_y^2\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2+\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{(\rho_{yz}-\rho_{xy}\rho_{yz})^2}{(1-\rho_{xz}^2)}\right] > 0 \tag{3.18}$$

where $D = \left[\left(\frac{C_x}{C_y}-\rho_{xy}\right)^2+\frac{(\rho_{yz}-\rho_{xy}\rho_{xz})^2}{(1-\rho_{xz}^2)}\right]$.

We note that the estimators (t_6, t_8) and (t_7, t_9) respectively are the members of the classes of estimators t_{10} and t_{11} and hence it can easily be shown that

$$\left.\begin{aligned}\min.\mathrm{MSE}(t_{10}) &< \mathrm{MSE}(t_6)\\ \min.\mathrm{MSE}(t_{10}) &< \mathrm{MSE}(t_8)\end{aligned}\right\} \tag{3.19}$$

and

$$\left.\begin{aligned}\min.\mathrm{MSE}(t_{11}) &< \mathrm{MSE}(t_7)\\ \min.\mathrm{MSE}(t_{11}) &< \mathrm{MSE}(t_9)\end{aligned}\right\} \tag{3.20}$$

Thus from (3.13) to (3.20), it follows that the proposed family of estimators 't' is more efficient than usual unbiased estimator $\bar{y}$, usual two-phase sampling ratio (product) estimator $t_1^{(R)}\left(t_1^{(P)}\right)$, Srivastava's (1970) estimator t_2, Chand's (1975) chain ratio estimator t_3, Srivastava et al.'s (1990) estimator t_4, Singh et al.'s (1994) estimator t_{15}, Singh and Upadhyaya's (1995) estimator t_5,

Upadhyaya and Singh's (2001) estimators $(t_6, t_7, t_8, t_9, t_{12}, t_{13}, t_{14})$ and the estimators $(t_{10},\ t_{11})$.

Finally, we conclude that the proposed family of estimators 't' is the best (in the sense of having minimum MSE).

Remark 3.1. We present a further generalized version of Singh et al.'s (1994) estimator t_{15} and the estimator t_{16} as

$$t^* = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_1}\left\{\frac{\psi(\phi_1(d))}{\psi(\phi_2(d))}\right\}^{\alpha_2}\left\{\frac{\psi^*(\phi_1(d))}{\psi^*(\phi_2(d))}\right\}^{\alpha_3} \tag{3.21}$$

where $\alpha_i's\ \ (i=1,2,3)$ are constants to be determined suitably,

$\psi(\phi_i(d)), (i=1,2)$ are same as defined for t_{15} at (3.1),

$$\psi^*(\phi_i(d)) = \phi_i(d) + \{1-\phi_i(d)\}\left(\frac{\bar{Z}}{z}\right),\ (i=1,2)$$

and $\phi_i(d),\ (i=1,2)$ are same as defined for t_{15} at (3.1).

A large number of estimators may be generated from t^* at (3.21) just by putting suitable values of $\alpha_i's\ \ (i=1,2,3)$ and d.

Expressions of bias and MSE of t^*, to the first order of approximation, are given by

$$\begin{aligned}\mathrm{B}(t^*) = \left(\frac{\bar{Y}}{2}\right)\Bigg[&\left(\frac{1}{n}-\frac{1}{N}\right)\phi^*\alpha_3\left\{\alpha_3\phi^* + 2K_{yz} - (\phi_1+\phi_2)\right\}C_z^2 \\ &+\left(\frac{1}{n'}-\frac{1}{N}\right)\phi^*\alpha_2\left\{\phi^*(\alpha_2+2\alpha_3) + 2K_{yz} - (\phi_1+\phi_2)\right\}C_Z^2 \\ &+\left(\frac{1}{n}-\frac{1}{n'}\right)\alpha_1\left\{(\alpha_1 - 2K_{yx}+1) - 2\alpha_3\phi^*K_{zx}\right\}C_x^2\Bigg]\end{aligned} \tag{3.22}$$

$$\begin{aligned}\mathrm{MSE}(t^*) = \bar{Y}^2\Bigg[&\left(\frac{1}{n}-\frac{1}{N}\right)\left\{C_y^2 + \phi^*\alpha_3 C_Z^2(\alpha_3\phi^* + 2K_{yz} - 2\alpha_1K_{xz})\right\} \\ &+\left(\frac{1}{n'}-\frac{1}{N}\right)\phi^*C_Z^2\left\{\alpha_2(\alpha_2\phi^* + 2K_{yz}) + 2\alpha_3(\alpha_2\phi^* + \alpha_1K_{xz})\right\} \\ &+\left(\frac{1}{n}-\frac{1}{n'}\right)\alpha_1C_x^2(\alpha_1 - 2K_{yx})\Bigg]\end{aligned} \tag{3.23}$$

where $\phi_i (i = 1,2)$ stands for $\phi_i (d)(i = 1,2)$; and $\phi^* = (\phi_1 - \phi_2)$.

The $\mathrm{MSE}(t^*)$ at (3.23) is minimized for

$$\left.\begin{aligned} \alpha_1 &= \frac{(K_{yx} - K_{yz}K_{zx})}{(1 - K_{xz}K_{zx})} = \alpha_{10}^* \quad \text{(say)} \\ \alpha_2 &= \frac{K_{xz}(K_{zx}K_{yz} - K_{yx})}{\phi^*(1 - K_{xz}K_{zx})} = \alpha_{20}^x \quad \text{(say)} \\ \alpha_3 &= \frac{(K_{zx}K_{yx} - K_{yz})}{\phi^*(1 - K_{xz}K_{zx})} = \alpha_{30}^* \quad \text{(say)} \end{aligned}\right\} \tag{3.24}$$

Thus the resulting (minimum) MSE of t^* is given by

$$\begin{aligned} \min.\mathrm{MSE}(t^*) &= S_y^2\left[\left(\frac{1}{n} - \frac{1}{N}\right) - \left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{yz}^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)R_{y.xz}^2\right] \\ &= \min.\mathrm{MSE(t)} \end{aligned} \tag{3.25}$$

where min. MSE(t) is same as given at (2.9).

Thus we state the following theorem.

Theorem 3.2. Up to terms of order n^{-1},

$$\min.\mathrm{MSE}(t^*) \geq S_y^2\left[\left(\frac{1}{n} - \frac{1}{N}\right) - \left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{yz}^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)R_{y.xz}^2\right]$$

with equality holding if

$$\alpha_1 = \alpha_{10}^*, \quad \alpha_2 = \alpha_{20}^*, \quad \alpha_3 = \alpha_{30}^*$$

where $\alpha_{i0}^{*}{}'s\ (i = 1,2,3)$ are defined in (3.24)

4. Estimators based on estimated optimum values of α_1, α_2 and α_3

The optimum values of α_1, α_2 and α_3 given at (2.7) can be expressed as

$$\alpha_{10} = \frac{\overline{X}}{\overline{Y}}\frac{(\beta_{yx} - \beta_{xz}\beta_{yz})}{(1 - \beta_{xz}\beta_{zx})},$$

$$\alpha_{20} = \frac{\overline{Z}}{\overline{Y}} \frac{\beta_{xz}(\beta_{yx} - \beta_{zx}\beta_{yz})}{(1-\beta_{xz}\beta_{zx})\phi} \ ,$$

$$\alpha_{30} = \frac{\overline{Z}}{\overline{Y}} \frac{(\beta_{yz} - \beta_{yx}\beta_{xz})}{(1-\beta_{xz}\beta_{zx})\phi}$$

where $\beta_{yx} = \frac{S_{yx}}{S_x^2}$, $\beta_{yz} = \frac{S_{yz}}{S_z^2}$, $\beta_{xz} = \frac{S_{xz}}{S_z^2}$, $\beta_{zx} = \frac{S_{zx}}{S_x^2}$.

The consistent estimates of α_{io} $(i = 1,2,3)$ based on sample data at hand are given by

$$\hat{\alpha}_{10} = \frac{\overline{x}}{\overline{y}} \frac{(\hat{\beta}_{yx} - \hat{\beta}_{xz}\hat{\beta}_{yz})}{(1-\hat{\beta}_{xz}\hat{\beta}_{zx})}$$

$$\hat{\alpha}_{20} = \frac{\overline{Z}}{\overline{y}} \frac{\hat{\beta}_{xz}(\hat{\beta}_{yx} - \hat{\beta}_{zx}\hat{\beta}_{yz})}{(1-\hat{\beta}_{xz}\hat{\beta}_{zx})\phi}, \qquad (\overline{Z} \text{ and } \phi \text{ are known})$$

$$\hat{\alpha}_{30} = \frac{\overline{Z}}{\overline{y}} \frac{(\hat{\beta}_{yz} - \hat{\beta}_{yx}\hat{\beta}_{xz})}{(1-\hat{\beta}_{xz}\hat{\beta}_{zx})\phi}, \quad (\overline{Z} \text{ and } \phi \text{ are known})$$

where $\hat{\beta}_{yx} = \frac{s_{yx}}{s_x^2}$, $\hat{\beta}_{xz} = \frac{s_{xz}}{s_z^2}$, $\hat{\beta}_{yz} = \frac{s_{yz}}{s_z^2}$, $\hat{\beta}_{zx} = \frac{s_{zx}}{s_x^2}$ are the consistent estimates of

β_{yx}, β_{xz}, β_{yz} and β_{zx} respectively,

$$s_{xy} = \sum_{i \in s_n} (x_i - \overline{x})(y_i - \overline{Y})/(n-1), \qquad s_{xz} = \sum_{i \in s_n} (x_i - \overline{x})(z_i - \overline{Z})/(n-1),$$

$$s_{yz} = \sum_{i \in s_n} (y_i - \overline{Y})(z_i - \overline{Z})/(n-1), \qquad s_x^2 = \sum_{i \in s_n} (x_i - \overline{X})^2/(n-1),$$

and $s_z^2 = \sum_{i \in s_n} (z_i - \overline{Z})^2/(n-1)$ are unbiased estimators of S_{xy}, S_{xz}, S_{yz}, S_x^2 and S_z^2 respectively.

Substituting $\hat{\alpha}_{io}$ $(i = 1,2,3)$ in place of α_i $(i = 1,2,3)$ in (2.1), we get the resulting estimator (based on estimated optimum values) for $\overline{Y}$ as

$$\hat{t} = \overline{y}\left(\frac{\overline{x}'}{\overline{x}}\right)^{\hat{\alpha}_{10}} \left(\frac{a\overline{Z}+b}{a\overline{z}'+b}\right)^{\hat{\alpha}_{20}} \left(\frac{a\overline{Z}+b}{a\overline{z}+b}\right)^{\hat{\alpha}_{30}}$$

$$= \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\hat{\alpha}_{10}}\left\{1+\phi\left(\frac{\bar{z}'-\bar{Z}}{\bar{Z}}\right)\right\}^{-\hat{\alpha}_{20}}\left\{1+\phi\left(\frac{\bar{z}-\bar{Z}}{\bar{Z}}\right)\right\}^{-\hat{\alpha}_{30}} \tag{4.1}$$

To obtain the MSE of $\hat{t}$, we write

$$\hat{\alpha}_{10} = \alpha_{10}(1+e_3), \quad \hat{\alpha}_{20} = \alpha_{20}(1+e_4), \quad \hat{\alpha}_{30} = \alpha_{30}(1+e_5) \text{ such that}$$

$$E(\hat{\alpha}_{10}) = \alpha_{10} + 0(n^{-1}) \quad \Rightarrow E(e_3) = 0(n^{-1})$$

$$E(\hat{\alpha}_{20}) = \alpha_{20} + 0(n^{-1}) \quad \Rightarrow E(e_4) = 0(n^{-1})$$

$$E(\hat{\alpha}_{30}) = \alpha_{30} + 0(n^{-1}) \quad \Rightarrow E(e_5) = 0(n^{-1}).$$

Thus expressing $\hat{t}$ in terms of $e's$, we have

$$\hat{t} = \bar{Y}(1+e_o)(1+e_1')^{\alpha_{10}(1+e_3)}(1+e_1)^{-\alpha_{10}(1+e_3)}(1+\phi e_2')^{-\alpha_{20}(1+e_4)}(1+\phi e_2)^{-\alpha_{30}(1+e_5)}$$
$$\cong \bar{Y}\left[1+e_o-\alpha_{10}(e_1-e_1')-\phi(\alpha_{20}e_2'+\alpha_{30}e_2)\right]$$

or

$$(\hat{t}-\bar{Y}) = \bar{Y}\left[e_o-\alpha_{10}(e_1-e_1')-\phi(\alpha_{20}e_2'+\alpha_{30}e_2\right] \tag{4.2}$$

Squaring both sides of the above expression and then taking expectation, we get the MSE of $\hat{t}$, to the first order of approximation, as

$$\text{MSE}(\hat{t}) = S_y^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)-\left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^2-\left(\frac{1}{n}-\frac{1}{n'}\right)R_{y.xz}^2\right]$$

$$= \text{min. MSE(t)} \tag{4.3}$$

Thus we proved the following theorem.

Theorem 4.1. If the optimum values of constants in the estimator t at (2.1) are replaced by their consistent estimators, the resulting estimator $\hat{t}$ has the same mean square error of 't' up to the first order of approximation, as that of optimum.

Further, let

$$\hat{\alpha}_{10}^* = \frac{\bar{x}}{\bar{y}}\frac{(\hat{\beta}_{yx}-\hat{\beta}_{xz}\hat{\beta}_{yz})}{(1-\hat{\beta}_{xz}\hat{\beta}_{zx})},$$

$$\hat{\alpha}_{20}^* = \frac{\bar{Z}}{\bar{y}}\frac{(\hat{\beta}_{zx}\hat{\beta}_{yz}-\hat{\beta}_{yx})\hat{\beta}_{xz}}{(1-\hat{\beta}_{xz}\hat{\beta}_{zx})\phi^*},$$

$$\hat{\alpha}_{30}^{*} = \frac{\overline{Z}}{\overline{y}} \frac{(\hat{\beta}_{yx}\hat{\beta}_{xz} - \hat{\beta}_{yz})}{(1-\hat{\beta}_{xz}\hat{\beta}_{zx})\phi^{*}}$$

be the consistent estimates of the optimum values α_{10}^{*}, α_{20}^{*} and α_{30}^{*} at (3.24) of α_1, α_2 and α_3 respectively.

Substituting $\hat{\alpha}_{i0}^{*}$ $(i=1,2,3)$ in place of $\alpha_i^{'}s$ $(i=1,2,3)$ in the estimator t^{*} at (3.21), we get the resulting estimator for $\overline{Y}$ as

$$\hat{t}^{*} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\hat{\alpha}_{10}^{*}} \left\{\frac{\psi(\phi_1(d))}{\psi(\phi_2(d))}\right\}^{\hat{\alpha}_{20}^{*}} \left\{\frac{\psi^{*}(\phi_1(d))}{\psi^{*}(\phi_2(d))}\right\}^{\hat{\alpha}_{30}^{*}}$$

$$= \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\hat{\alpha}_{10}^{*}} \left\{\frac{\left(1+\phi_1(d)\frac{(\bar{z}'-\bar{Z})}{\bar{Z}}\right)}{\left(1+\phi_2(d)\left(\frac{\bar{z}'-\bar{Z}}{\bar{Z}}\right)\right)}\right\}^{\hat{\alpha}_{20}^{*}} \left\{\frac{\left(1+\phi_1(d)\left(\frac{\bar{z}-\bar{Z}}{\bar{Z}}\right)\right)}{\left(1+\phi_2(d)\left(\frac{\bar{z}-\bar{Z}}{\bar{Z}}\right)\right)}\right\}^{\hat{\alpha}_{30}^{*}} \tag{4.4}$$

To obtain the MSE of $\hat{t}^{*}$, we write

$$\hat{\alpha}_{10}^{*} = \alpha_{10}^{*}(1+e_3^{*}), \qquad \hat{\alpha}_{20}^{*} = \alpha_{20}^{*}(1+e_4^{*}),$$

$$\hat{\alpha}_{30}^{*} = \alpha_{30}^{*}(1+e_5^{*})$$

such that $E(\hat{\alpha}_{10}^{*}) = \alpha_{10}^{*} + 0(n^{-1}) \Rightarrow \mathrm{E}(e_3^{*}) = 0(n^{-1})$,

$E(\hat{\alpha}_{20}^{*}) = \alpha_{20}^{*} + 0(n^{-1}) \Rightarrow \mathrm{E}(e_4^{*}) = 0(n^{-1})$,

$E(\hat{\alpha}_{30}^{*}) = \alpha_{30}^{*} + 0(n^{-1}) \Rightarrow \mathrm{E}(e_5^{*}) = 0(n^{-1})$.

Expressing (4.4) in terms $e's$ and $e^{*}'s$, we have

$$\hat{t}^{*} = \overline{Y}(1+e_o)\left\{\frac{(1+e_1')}{(1+e_1)}\right\}^{\alpha_{10}^{*}(1+e_3^{*})} \left\{\frac{1+\phi_1(d)e_2'}{1+\phi_2(d)e_2'}\right\}^{\alpha_{20}^{*}(1+e_4^{*})} \left\{\frac{1+\phi_1(d)e_2}{1+\phi_2(d)e_2}\right\}^{\alpha_{30}^{*}(1+e_5^{*})}$$

or

$$\left(\hat{t}^{*} - \overline{Y}\right) \cong \overline{Y}\left[e_0 - \alpha_{10}^{*}\left(e_1 - e_1'\right) + \alpha_{20}^{*}\phi^{*}e_2' + \alpha_{30}^{*}\phi^{*}e_2\right]$$

Squaring both sides and then taking expectation, we get the MSE of $\hat{t}^{*}$, to the first order of approximation, as

$$\mathrm{MSE}(\hat{t}^{*}) = \overline{Y}^{2}\left[\left(\frac{1}{n}-\frac{1}{N}\right) - \left(\frac{1}{n'}-\frac{1}{N}\right)\rho_{yz}^{2} - \left(\frac{1}{n}-\frac{1}{n'}\right)R_{y.xz}^{2}\right]$$

$$= \min.\mathrm{MSE}(t^*)$$

Thus we proved the following theorem.

Theorem 4.2. If the optimum values of constants in the estimator t^* at (3.21) are replaced by their consistent estimators, the resulting estimator $\hat{t}^*$ has the same mean square error, up to the first order of approximation, as that of optimum t^*.

5. Empirical Study

The various results obtained in previous sections are now examined with the help of the data used earlier by Anderson (1958). Twenty five 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, σ_z = 7.3747,

C_y = 0.0546, C_x= 0.0526, C_z = 0.0488, $\beta_1(z)$ = 0.0002, $\beta_2(z)$ = 2.6519,

ρ_{xy} = 0.7108, ρ_{yz} = 0.6932, ρ_{xz} = 0.7346.

Take n = 7 and n' =10.

We have computed the percent relative efficiency (PRE) of different estimators of $\bar{Y}$ with respect to usual unbiased estimator $\bar{y}$ and compiled in Table 4.1.

Table 4.1. Percent relative efficiencies of different estimators of $\bar{Y}$ with respect to $\bar{y}$.

Estimator	MSE/min. MSE /Variance	PRE (., $\bar{y}$)	Estimator	MSE/min. MSE /Variance	PRE (., $\bar{y}$)
$\bar{y}$	10.3633	100.00	t_6	5.7959	178.80
$t_1^{(R)}$	8.4573	122.54	t_7	5.5617	186.33
$t_1^{(P)}$	10.364	99.99	t_8	20.7753	49.88
t_2	8.1818	126.66	t_9	16.5795	62.51
t_3	5.7956	178.81	$t_j\,(j = 10 \text{ to } 15)$	5.5524	186.65
t_4	5.2769	196.39	t_{16}	5.2769	196.39
t_5	5.5524	186.65	t (or $\hat{t}$) and t^* ($\hat{t}^*$)	5.0025	207.16

From Table 4.1 we can make the following observations:

(i) The estimator t_2 considered by Srivastava (1971) at its optimum condition is more efficient than the usual unbiased estimator $\bar{y}$, usual two phase sampling ratio $t_1^{(R)}$ and product $t_1^{(P)}$ estimators. Even product estimator $t_1^{(P)}$ is inferior to usual unbiased estimator $\bar{y}$ which does not make use of information on any auxiliary variable. Thus the use of the estimator t_2 is recommended in practice when information on only single (main) auxiliary variable is available.

(ii) The estimators t_5 and $t_j\,(j = 10 \text{ to } 15)$ respectively reported by Singh and Upadhyaya (1996) and Singh (2001) are equally efficient at their optimum conditions. Also the estimator t_7 envisaged by Upadhyaya and Singh (2001) has efficiency almost equal to that of the estimators t_5 and $t_j\,(j = 10 \text{ to } 15)$.

(iii) Chand's (1975) chain estimator t_3 and Upadhyaya and Singh's (2001) estimator t_6 are almost equally efficient but inferior to the estimators t_5, t_7 and $t_j\,(j = 10 \text{ to } 15)$. This leads that the estimator t_6 which uses additional information on coefficient of variation and the coefficient of kurtosis of second auxiliary variable, has no improvement over Chand's (1975) chain estimator t_3. Thus the estimator t_3 is to be preferred over t_6.

(iv) Owing to the positive correlation between the study variable y and the main auxiliary variable x, the usual two phase sampling product estimator $t_1^{(P)}$ is inferior to the sample mean $\bar{y}$. The estimators t_8 and t_9 are also inferior to the estimators $t_1^{(P)}$ and $\bar{y}$ due to the positive correlation between (y and x) and (x and z). Thus we note that the estimators $t_1^{(P)}$, t_8 and t_9 are not suitable to the considered data set .

(v) The estimator t_4 considered by Srivastava et al (1990) and the estimator t_{16} are equally efficient at their optimum conditions which are superior to the estimators $\bar{y}$, $t_1^{(R)}$, $t_1^{(P)}$, t_2, t_3 and $t_j\,(j = 10 \text{ to } 15)$.

(vi) The proposed estimators t ($\hat{t}$) and $t^*(\hat{t}^*)$ are better than the estimators t_4 and t_{16} with substantial gain in efficiency and hence with all the estimators discussed here.

Finally we conclude that the proposed estimators t ($\hat{t}$) and $t^*(\hat{t}^*)$, in the presence of two auxiliary variable, are most reliable for researchers.

REFERENCES

AHMED, M.S (1998), A note on regression-type estimators using multiple auxiliary information. Australian & New Zealand Journal of Statistics. 40(3), 373—376.

AHMED, M.S., KHAN, S.U. and TRIPATHI, T. P. (1994), Two general class of chain ratio and product estimators for a finite population mean based on two-phase sampling and multivariate information. Jour. Stat. Stud., 14, 86—99.

ANDERSON, T. W. (1958), An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Inc., New York.

BEDI, P. K. (1985), On two-phase multivariate sampling estimator. Journal of Indian Society of Agricultural Statistics. 37(2), 158—162.

CHAND, L. (1975), Some ratio type estimators based on two or more auxiliary variables. Unpublished Ph. D. thesis, Iowa State University, Ames, Iowa (USA).

GUPTA, R. K., SINGH, S. and MANGAT, N. S. (1992—93), Some chain ratio type estimators for estimating finite population variance. Aligarh Journal of Statistics., 12 & 13, 65—69.

HIDIROGLOU, M. A. and SARNDAL, C.E. (1995), Use of auxiliary information for two-phase sampling. Proceedings of the section on Survey Research Methods, American Statistical Association, Vol.II, 873—878.

HIDIROGLOU, M.. A. and SARNDAL, C.E. (1998), Use of auxiliary information for two — phase sampling. Survey Methodology, 24(1), 11—20.

KIREGYERA, B. (1980), A chain ratio type estimators in finite population double sampling using two auxiliary variables. Metrika, 17, 217—223.

KIREGYERA, B. (1984), Regression type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrika, 31, 215—226.

LUI K.J. (1990), Modified product estimators of finite population mean in finite sampling. Communications in Statistics, Theory and Methods, 19(10), 3799—3807.

MURTHY, M..N (1967), Sampling Theory and Methods. Statistical Publishing Society, Calcutta, India.

MUKERJEE, R., RAO, T. J. and VIJAYAN, K. (1987), Regression type estimators using multiple auxiliary information. Australian Journal of Statistics, 29(3), 244—254.

NAIK, V. D. and GUPTA, P. C. (1991), A general class of estimators for estimating population mean using auxiliary information. Metrika, 38, 11—17.

PRASAD, B. (1989), Some improved ratio-type estimators of population mean and ratio in finite population sample surveys. Communications in Statistics, Theory and Methods, 18, 379—392.

PRASAD, B., SINGH, R. S. and SINGH, H. P. (1996), Some chain ratio-type estimators for ratio of two population means using two auxiliary characters in two phase sampling. Metron, LIV, (1—2), 95—113.

REDDY, V. N. (1993), On ratio and product methods of estimation. Sankhya, B, 35, 307—316.

SAHOO, J. and SAHOO, L. N. (1993), A class of estimators in two-phase sampling using two auxiliary variables. Journal of Indian Society of Agricultural Statistics, 31, 107—114.

SAHOO, J. and SAHOO, L. N. (1994), On the efficiency of four chain-type estimators in two-phase sampling under a model. Statistics, 25, 361—366.

SAHOO, J. and SAHOO, L. N. (1999 a), An alternative class of estimators in double sampling procedures. Calcutta Statistical Association Bulletin, 49, 79—83.

SAHOO, J. and SAHOO, L. N. (1999 b), A comparative study of some regression-type estimators in double sampling procedures. Aligarh Journal of Statistics, 19, 67—76.

SAHOO, J., SAHOO, L. N. and MOHANTY, S. (1994 a), A regression approach to estimation in two-phase sampling using two auxiliary variables. Current Science, 65(1), 73—75.

SAHOO, J., SAHOO, L. N. and MOHANTY, S. (1994 b), An alterntive approach to estimation in two-phase sampling using two auxiliary variables. Biometrical Journal 3, 293—298.

SAHOO, L. N. and SAHOO, R. K. (2001), Predictive estimation of finite population mean in two-phase sampling using two auxiliary variables. Journal of Indian Society of Agricultural Statistics, 54(2), 258—264.

SINGH, G. N. (2001), On the use of transformed auxiliary variable in the estimation of population mean in two phase sampling. Statistics in Transition, 5(3), 405—416.

SINGH, G. N. and UPADHYAYA, L. N. (1995), A class of modified chain type estimators using two auxiliary variables in two phase sampling. Metron, LIII, 117—125.

SINGH, H. P. (1993), A chain ratio-cum-difference estimator using two auxiliary variates in double sampling. Journal of Ravishankar University, 6, B, (Science), 79—83.

SINGH, H. P. and ESPEJO, M. R. (2000), An improved class of chain regression estimators in two phase sampling. Statistics & Decisions 18, 205—218.

SINGH, H. P. and GANGELE, R. K. (1989-95), An improved regression estimator with known coefficient of variation in two-phase sampling with two auxiliary variables. The Vikram, 25 (1 & 3), 27—40.

SINGH, HOUSILA, P., SINGH, HARI, P. and KUSHWAHA, K. S. (1992), On chain ratio-to-regression-type estimator in double sampling. Assam Statistical Review, 6(2), 91—105.

SINGH, V. K. and SINGH, G. N. (1991), Chain type regression estimators with two auxiliary variables under double sample scheme. Metron, XL1X, (1—4), 279—289.

SINGH, V. K. SINGH, G. N. and SHUKLA, D. (1994), A class of chain ratio type estimators with two auxiliary variables under double sampling scheme. Sankhya, 56, B, 209—221.

SINGH, V. K., SINGH, HARI P., SINGH, HOUSILA P. and SHUKLA, D. (1994), A general class of chain estimators for ratio and product of two means of a finite population. Communications in Statistics, Theory and Methods., 2395), 1341—1355.

SINGH, V. P. and SINGH, H. P. (1997-98), Chain estimators for population ratio in double sampling, Aligarh Journal of Statistics, (17 & 18), 85—100.

SRIVASTAVA, S. K. (1967), An estimator using auxiliary information in sample surveys, Calcutta Statistical Association Bulletin, 16, 121—132.

SRIVASTAVA, S. K. (1970), A two-phase sampling estimator in sample surveys. Australian Journal of Statistics, 12(1) 23—27.

SRIVASTAVA, S. K. (1981). A generalized two-phase sampling estimator. Journal of Indian Society of Agricultural Statistics, 33(1), 38—46.

SRIVASTAVA, RANI S., KHARE, B. B. and SRIVASTAVA, S. R. (1988), On generalized chain estimator for ratio and product of two population means using auxiliary characters. Assam Statistical Review, 2, 21—29.

SRIVASTAVA, RANI S., SRIVASTAVA, S. R. and KHARE, B. B. (1989), Chain ratio type estimator for ratio of two population means using auxiliary

characters. Communications in Statistics, Theory and Methods, 18(10), 3917—3926.

SRIVASTAVA, RANI S., KHARE, B. B. and SRIVASTAVA, S. R. (1990), A generalized chain ratio estimator for mean of finite population. Journal of Indian Society of Agricultural Statistics, 42(1), 108—117.

SRIVENKATARAMANA, T. and TRACY, D. S. (1989), Two-phase sampling for selection with probability proportional to size in sample surveys. Biometrika, 76, 4, 818—821.

TRACY, D. S. and SINGH, H. P.(1999), A general class of chain regression estimators in two phase sampling. Journal of Applied Statistical Science, 8, (4), 205—216.

TRIPATHI, T. P. and AHMED, M. S. (1995), A class of estimators for a finite population mean based on multivariate information and general two-phase sampling. Calcutta Statistical Association Bulletin, 45, 203—218.

UPADHYAYA, L. N., KUSHWAHA, K. S. and SINGH, H. P. (1990), A modified chain ratio-type estimator in two phase sampling using multi auxiliary information. Metron, 48, (1—4), 381—393.

UPADHYAYA, L. N., DUBEY, S. P. and SINGH, H. P. (1992), A class of ratio-in-regression estimators using two auxiliary variables in double sampling. The Journal of Scientific Research, Banaras Hindu University, Varanasi, 42, 127—134.

UPADHYAYA, L. N. and SINGH, G. N. (2001), Chain type estimators using transformed auxiliary variable in two-phase sampling. Advances in Modeling and Analysis, 38, (1—2), 1—10.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1079—1093

ESTIMATION OF POPULATION MEAN WITH KNOWN COEFFICIENT OF VARIATION UNDER OPTIONAL RANDOMIZED RESPONSE MODEL USING SCRAMBLED RESPONSE TECHNIQUE

Housila P. Singh and Nidhi Mathur[1]

ABSTRACT

In this paper we have suggested an estimator of population mean μ_x when the coefficient of variation C_x is available; and its properties are studied. An estimator for μ_x is also suggested when a prior value μ_{x_0} of μ_x is available along with its properties. Numerical illustrations are given to judge the merits of the suggested estimators over Gupta et. al. (2002) estimator.

Key words: Optional randomized response, Scrambled randomized response, Prior estimate.

1. Introduction

The randomized response technique (RRT), an ingenious interviewing procedure for eliciting information on sensitive characters was introduced by Warner (1965) and it has been extensively studied by various authors. Eichhorn and Hayre (1983) suggested a method known as the scrambled randomized response method. In this method, "the respondent multiplying his answers to the sensitive question by a random number from a known distribution. The product is given to the interviewer, who does not know the value of the random number".

Let X denote the answer to the sensitive question, and let S be a random variable independent of X and with finite mean and variance. It is assumed that $X \geq 0$ and $S>0$. The respondent generates S using some specified method and multiplies his sensitive answer X by S. The interviewer receives the scrambled answer $Y=XS$. The particular values of S are unknown to the interviewer, but its

[1] School of Studies in Statistics, Vikram University, Ujjain—456010, India.
e-mail:hpsujn@rediffmail.com/nidhimathur_29@yahoo.co.in

distribution is known. The mean θ =E(*S*) and the variance γ^2 =V(*S*) of the scrambling variables *S*, are assumed to be known.

The i^{th} respondent selected in the sample of size *n*, drawn by using SRSWR, is requested to report the value $z_j = y_j/\theta$ as a scrambled response on the sensitive variable, *X*. An unbiased estimator of population mean, $\mu_x = \mathrm{E}(X)$ is given by

$$\hat{\mu}_s = \frac{1}{n}\sum_{j=1}^{n} z_j = \bar{z} \tag{1.1}$$

which is due to Eichhorn and Hayre (1983).
The variance of $\hat{\mu}_s$ is given by

$$\mathrm{V}(\hat{\mu}_s) = \frac{\mu_x^2 \mathrm{V}_x}{n} \tag{1.2}$$

where $\sigma_x^2 = \mathrm{V}(X)$, $C_x = \sigma_x/\mu_x$, $\mathrm{V}_x = \left[C_x^2 + C_\gamma^2\left(1+C_x^2\right)\right]$, and $C_\gamma = \gamma/\theta$ denotes the known coefficient of variation of the scrambling variable *S*.

Singh and Mathur (2003) suggested a class of estimators for μ_x as

$$\hat{\mu}_M = M\ \bar{z}\ , \tag{1.3}$$

where *M* is a suitably chosen constant such that $0<M\leq 1$.

The bias and mean square error (MSE) of $\hat{\mu}_M$ are respectively given by

$$\mathrm{B}(\hat{\mu}_M) = (M-1)\mu_x \tag{1.4}$$

and

$$\mathrm{MSE}(\hat{\mu}_M) = \mu_x^2\left[M^2\left(1+\frac{\mathrm{V}_x}{n}\right) - 2M + 1\right] \tag{1.5}$$

The minimization of MSE($\hat{\mu}_M$) with respect to M yields the optimum value of M as

$$M = \frac{n}{(n+\mathrm{V}_x)} = M_{opt}\,(\text{say}) \tag{1.6}$$

This leads to the optimum estimator in the class $\hat{\mu}_s$ as

$$\hat{\mu}_{M_0} = \frac{n}{(n+\mathrm{V}_x)}\bar{z} \tag{1.7}$$

The bias and mean square error (MSE) of $\hat{\mu}_{M_0}$ are respectively given by

$$\mathrm{B}\left(\hat{\mu}_{M_0}\right) = -\frac{\mathrm{V}_x\,\mu_x}{\left(n+\mathrm{V}_x\right)} \tag{1.8}$$

and $$\mathrm{MSE}\left(\hat{\mu}_{M_0}\right) = \frac{\mathrm{V}_x\,\mu_x^2}{\left(n+\mathrm{V}_x\right)} \tag{1.9}$$

It is obvious from (1.8) and (1.9) that

$$\frac{\mathrm{MSE}\left(\hat{\mu}_{M_0}\right)}{\mu_x^2} = \frac{\left|\mathrm{B}\left(\hat{\mu}_{M_0}\right)\right|}{\mu_x} \tag{1.10}$$

i.e. $\mathrm{RMSE}(\hat{\mu}_{M_0}) = \left|\mathrm{RB}\left(\hat{\mu}_{M_0}\right)\right|$

From (1.2) and (1.9) we note that $\hat{\mu}_{M_0}$ is more efficient than $\hat{\mu}_s$ if

$\mathrm{MSE}(\hat{\mu}_{M_0}) < \mathrm{V}(\hat{\mu}_s)$

i.e. if $$\frac{\mathrm{V}_x\,\mu_x^2}{\left(n+\mathrm{V}_x\right)} < \frac{\mathrm{V}_x\,\mu_x^2}{n}$$

i.e. if $$\frac{1}{\left(n+\mathrm{V}_x\right)} < \frac{1}{n} \quad , \tag{1.11}$$

which is always true. Thus the optimum estimator $\hat{\mu}_{M_0}$ is always better than the unbiased estimator $\hat{\mu}_s$ when the coefficient of variation (CV) C_x of X is known.

In this paper we have suggested an estimator of μ_x when C_x of X is known under optional randomized response technique and its properties are studied. Numerical illustrations are given to judge the merits of the suggested estimators over Gupta et. al. (2002)estimators.

2. The Suggested Estimator

Gupta et. al. (2002) have suggested an alternative randomized response technique, which is more efficient than the one suggested by Eichhorn and Hayre (1983) model. Although it is not a serious drawback of Eichhorn and Hayre (1983) model, but some respondents may not like to multiply or may not know how to multiply. Moreover, depending upon the nature of the question, some respondents may not feel any need to scramble the response. Still, we hesitate to ask directly, because some respondents may give untruthful response, which is more dangerous than the scrambled response, as illustrated by Singh et. al. (1996). This led Gupta et. al. (2002) to suggest an optional randomization technique described below:

In optional randomized response technique, each respondent selected in the SRSWR sample chooses one of the following two options:
(a) The respondent can report the correct response *X*, or
(b) The respondent can report the scrambled response *SX*,

where *S* denotes the independent scrambling variable described earlier.

In this procedure it is assumed that both *S* and *X* are positive valued random variables and that $\mu_s = \mathrm{E}(S) = \theta(=1)$ and $\sigma_s^2 = \mathrm{V}(S) = \gamma^2$. The optional randomized response model then can be written as

$$U = S^{\delta} X\,, \tag{2.1}$$

where δ is a random variable defined as

$$\delta = \begin{cases} 1, \text{ if the response is scrambled,} \\ 0, \text{ otherwise.} \end{cases} \tag{2.2}$$

If *w* denotes the probability that a person will report the scrambled response, then δ is a Bernouill variate with E(δ)= *w* known as the sensitivity level of the question. If a question in the survey is more sensitive then more people will report scrambled responses and the value of *w* will be near to 1. If the question is not very sensitive, then the value of *w* will be close to '0'. *W* is also termed as a measure of the level of sensitivity of the question in the personal interview surveys.

An unbiased estimator of the population mean μ_x is given by

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} u_i \tag{2.3}$$

which due to Gupta et. al. (2002).

The variance of $\hat{\mu}_x$ is given by

$$\mathrm{V}(\hat{\mu}_x) = \frac{\mu_x^2\, \mathrm{V}_{wx}}{n}, \tag{2.4}$$

where $\mathrm{V}_{wx} = \left[C_x^2 + wC_\gamma^2\left(1 + C_x^2\right)\right]$.

Motivated by Searls (1964) and Searls and Intarapanich (1990) we suggest a family of estimators of μ_x as

$$\hat{\mu}_{A_1} = A_1 \hat{\mu}_x\,, \tag{2.5}$$

A_1 being a suitably chosen scalar. The family $\hat{\mu}_{A_1}$ reduces to the following set of estimators of μ_x as

(i) $\hat{\mu}_{A_1} \to \hat{\mu}_s$ for A_1 =1, δ =1,

(ii) $\hat{\mu}_{A_1} \to \hat{\mu}_M$ for A_1 =*M*, δ =1,

(iii) $\hat{\mu}_{A_1} \to \hat{\mu}_x$ for A_1 =1.

The bias and mean square error (MSE) of $\hat{\mu}_{A_1}$ are respectively given by

$$\mathrm{B}(\hat{\mu}_{A_1}) = (A_1 - 1)\mu_x \tag{2.6}$$

and $$\mathrm{MSE}(\hat{\mu}_{A_1}) = A_1^2\, \mathrm{V}(\hat{\mu}_x) + \left(\mathrm{B}(\hat{\mu}_{A_1})\right)^2$$

$$= \mu_x^2 \left[A_1^2 \left(1 + \frac{\mathrm{V}_{wx}}{n} \right) - 2A_1 + 1 \right] \tag{2.7}$$

From (2.4) and (2.7) we note that $\mathrm{MSE}(\hat{\mu}_{A_1}) < \mathrm{V}(\hat{\mu}_x)$ if

$$\left(\frac{n - \mathrm{V}_{wx}}{n + \mathrm{V}_{wx}} \right) < A_1 < 1 \tag{2.8}$$

The range of A_1 have been computed for different values of *n*, C_γ, C_x for which the proposed estimator $\hat{\mu}_{A_1}$ is more efficient than $\hat{\mu}_x$ and shown in Table-2.1.

The MSE of $\hat{\mu}_{A_1}$ at (2.7) is minimum when

$$A_1 = \frac{n}{(n + \mathrm{V}_{wx})} \tag{2.9}$$

Thus the resulting optimum estimator in the class $\hat{\mu}_{A_1}$ is given by

$$\hat{\mu}_{A_{10}} = \frac{n}{(n + \mathrm{V}_{wx})} \hat{\mu}_x \tag{2.10}$$

which requires the prior knowledge of the coefficient of variation C_x of *X* as C_γ is known.

The bias and mean square error(MSE) of $\hat{\mu}_{A_{10}}$ are respectively given by

$$\mathrm{B}(\hat{\mu}_{A_{10}}) = -\frac{\mu_x \mathrm{V}_{wx}}{(n + \mathrm{V}_{wx})} \tag{2.11}$$

and $$\text{MSE}(\hat{\mu}_{A_{10}})=\frac{\mu_x^2 \text{V}_{wx}}{(n+\text{V}_{wx})} \tag{2.12}$$

It follows from (2.11) and (2.12) that

$$\text{RMSE}(\hat{\mu}_{A_{10}})=\left|\text{RB}(\hat{\mu}_{A_{10}})\right| \tag{2.13}$$

From (2.4) and (2.12) w have

$$\text{MSE}(\hat{\mu}_x)-\text{MSE}(\hat{\mu}_{A_{10}})=\frac{\mu_x^2 \text{V}_{wx}^2}{n(n+\text{V}_{wx})}>0 \tag{2.14}$$

which indicates that the optimum estimator $\hat{\mu}_{A_{10}}$ is always better than usual unbiased estimator $\hat{\mu}_x$ when the coefficient of variation C_x of X is known.

Further, from (1.2), (1.9), (2.4), (2.12) we have

$$\text{V}(\hat{\mu}_s)-\text{V}(\hat{\mu}_x)=(\mu_x^2/n)C_\gamma^2(1+C_x^2)(1-w)\geq 0 \tag{2.15}$$

$$\text{MSE}(\hat{\mu}_{M_0})-\text{MSE}(\hat{\mu}_{A_{10}})=\frac{n\mu_x^2 C_\gamma^2(1+C_x^2)(1-w)}{(n+\text{V}_x)(n+\text{V}_{wx})}\geq 0 \tag{2.16}$$

since $0<w\leq 1$.

Thus from (1.11), (2.14), (2.15) and (2.16) we have the following inequalities:

$$\text{MSE}(\hat{\mu}_{A_{10}})<\text{V}(\hat{\mu}_x)<\text{V}(\hat{\mu}_s) \tag{2.17}$$

and $$\text{MSE}(\hat{\mu}_{A_{10}})<\text{MSE}(\hat{\mu}_{M_0})<\text{V}(\hat{\mu}_s) \tag{2.18}$$

since $0<w\leq 1$.

which shows that the proposed estimator $\hat{\mu}_{A_{10}}$ is more efficient than Eichhorn and Hayre's (1983) estimator $\hat{\mu}_s$,Gupta et. al.'s (2002) estimator $\hat{\mu}_x$ and Singh and Mathur (2003) $\hat{\mu}_{M_0}$, when the coefficient of variation C_x of X is known.

Table-2.1. Range of A_1 for different values of n, w, C_γ and C_x.

C_γ **=1.0, *w*=0.1**					
$C_x \downarrow$ ***n*** $\rightarrow$	**5**	**10**	**15**	**20**	**50**
0.25	0.93~1	0.97~1	0.98~1	0.98~1	0.99~1
0.5	0.86~1	0.93~1	0.95~1	0.96~1	0.99~1
0.75	0.75~1	0.87~1	0.91~1	0.93~1	0.97~1
1	0.61~1	0.79~1	0.85~1	0.89~1	0.95~1
2	0.05~1	0.38~1	0.54~1	0.63~1	0.83~1
3	0~1	0~1	0.20~1	0.33~1	0.67~1
4	0~1	0~1	0~1	0.06~1	0.48~1
5	0~1	0~1	0~1	0~1	0.29~1
6	0~1	0~1	0~1	0~1	0.11~1
***w*=0.3**					
0.25	0.86~1	0.93~1	0.95~1	0.96~1	0.98~1
0.5	0.78~1	0.88~1	0.92~1	0.94~1	0.98~1
0.75	0.66~1	0.81~1	0.87~1	0.90~1	0.96~1
1	0.52~1	0.72~1	0.81~1	0.85~1	0.94~1
2	0~1	0.29~1	0.46~1	0.57~1	0.80~1
3	0~1	0~1	0.11~1	0.25~1	0.61~1
4	0~1	0~1	0~1	0~1	0.41~1
5	0~1	0~1	0~1	0~1	0.21~1
6	0~1	0~1	0~1	0~1	0.03~1
***w*=0.5**					
0.25	0.79~1	0.89~1	0.92~1	0.94~1	0.98~1
0.5	0.70~1	0.84~1	0.89~1	0.92~1	0.97~1
0.75	0.58~1	0.76~1	0.84~1	0.87~1	0.95~1
1	0.43~1	0.67~1	0.76~1	0.82~1	0.92~1
2	0~1	0.21~1	0.40~1	0.51~1	0.77~1
3	0~1	0~1	0.03~1	0.18~1	0.56~1
4	0~1	0~1	0~1	0~1	0.34~1
5	0~1	0~1	0~1	0~1	0.14~1
6	0~1	0~1	0~1	0~1	0~1
***w*=0.9**					
0.25	0.66~1	0.82~1	0.87~1	0.90~1	0.96~1
0.5	0.57~1	0.76~1	0.83~1	0.87~1	0.95~1
0.75	0.43~1	0.67~1	0.77~1	0.82~1	0.92~1
1	0.28~1	0.56~1	0.69~1	0.75~1	0.89~1
2	0~1	0.08~1	0.28~1	0.40~1	0.71~1
3	0~1	0~1	0~1	0.05~1	0.47~1
4	0~1	0~1	0~1	0~1	0.23~1
5	0~1	0~1	0~1	0~1	0.02~1
6	0~1	0~1	0~1	0~1	0~1

Table 2.1 demonstrates the range of A_1 for different values of C_x,*n*, *w* and C_γ=1. We find that the estimator $\hat{\mu}_{A_1}$ is more efficient than Gupta et. al. (2002) estimator $\hat{\mu}_x$ for wider range of A_1 when C_x is large and *n* is small. The magnitude of the range of A_1 increases as *w* increases (i.e. sensitivity level of the question increases).

3. Estimator Based On Estimated (Or Guessed) Value Of C_x

Sometimes, it is possible that the experimenters may not have the accurate information about C_x. They may have the guess value $C_{x_0} = aC_x (a > 0)$. In this case the estimator of μ_x is

$$\hat{\mu}^*_{A_{10}} = \frac{n}{\left(n + \mathrm{V}^{(0)}_{wx}\right)} \hat{\mu}_x , \tag{3.1}$$

where $\mathrm{V}^{(0)}_{wx} = \left[C^2_{x_0} + wC^2_\gamma \left(1 + C^2_{x_0}\right)\right]$.

The bias and mean square error (MSE) of $\hat{\mu}^*_{A_{10}}$ are respectively given by

$$\mathrm{B}(\hat{\mu}^*_{A_{10}}) = -\frac{\mathrm{V}^{(0)}_{wx}\,\mu_x}{\left(n + \mathrm{V}^{(0)}_{wx}\right)} \tag{3.2}$$

and $$\mathrm{MSE}(\hat{\mu}^*_{A_{10}}) = \frac{\mu^2_x \left(n\,\mathrm{V}_{wx} + \mathrm{V}^{(0)2}_{wx}\right)}{\left(n + \mathrm{V}^{(0)}_{wx}\right)^2} . \tag{3.3}$$

The estimator $\hat{\mu}^*_{A_{10}}$ will be more efficient than $\hat{\mu}_x$ if the following inequality is satisfied:

$$a^2 \le \frac{\left[\left(2n + C^2_\gamma\right)\mathrm{V}_{wx} - C^2_\gamma \left\{nw + C^2_x \left(1 + wC^2_\gamma\right)(1 - w)\right\}\right]}{C^2_x \left(1 + wC^2_\gamma\right)\left[n - \mathrm{V}_{wx} - (1 - w)C^2_\gamma\right]} \tag{3.4}$$

For given *n*, C_x, C_γ and *w*, one can calculate the range of *a*.

4. Shrinkage Estimator When a Guessed Value μ_{x_0} of μ_x Is Known

Following Thompson (1968) we suggest a shrinkage estimator for μ_x as

$$\hat{\mu}^*_k = k^* \hat{\mu}_x + \left(1 - k^*\right)\mu_{x_0}, 0 < k^* \le 1 \tag{4.1}$$

where μ_{x_0} is a guessed value of μ_x.

The bias and mean square error (MSE) of $\hat{\mu}_k^*$ are respectively given by

$$\mathrm{B}(\hat{\mu}_k^*) = \left(1-k^*\right)\mu_x\Delta \tag{4.2}$$

and

$$\mathrm{MSE}(\hat{\mu}_k^*) = \mu_x^2\left[\frac{k^{*2}\,\mathrm{V}_{wx}}{n} + \left(1-k^*\right)^2\ \Delta^2\right] \tag{4.3}$$

where $\Delta = \left(\mu_{x_0} - \mu_x\right)\big/\mu_x$.

It follows from (2.4) and (4.3) that $\mathrm{MSE}(\hat{\mu}_k^*) < \mathrm{V}(\hat{\mu}_x)$ if

$$\left(\frac{n\Delta^2 - \mathrm{V}_{wx}}{n\Delta^2 + \mathrm{V}_{wx}}\right) < k^* \le 1 \tag{4.4}$$

For given C_x, C_γ, *w*, *n* and Δ, we have calculated the ranges of k^* for which the proposed estimator $\hat{\mu}_k^*$ is more efficient than $\hat{\mu}_x$ in Table-4.1.

Table 4.1, presents the range of k^* for different values of *n*, Δ, *w* and C_γ =1. We observe that for $0.1 \le w \le 0.9$ and Δ= 0.1,0.3; the estimator $\hat{\mu}_k^*$ is more efficient than Gupta et. al.'s (2002) estimator $\hat{\mu}_x$ for full range of k^* when $C_x\ (\ge 2)$ and all *n*. For $0.1 \le w \le 0.9$ and Δ= 0.5,0.9, the full range of k^* is obtained when $C_x \ge 4$ and $\mathrm{n} \le 20$. Thus we note that the estimator $\hat{\mu}_k^*$ is to be used in practice when the population is heterogenous.

Table-4.1. Range of k^* for different values of n, C_x, C_γ, w and Δ.

C_γ =1.0, *w*=0.1										
Δ=0.1						**Δ=0.3**				
$n\rightarrow$ $C_x\downarrow$	**5**	**10**	**15**	**20**	**50**	**5**	**10**	**15**	**20**	**50**
0.25	0~1	0~1	0~1	0.08~1	0.50~1	0.45~1	0.68~1	0.78~1	0.83~1	0.93~1
0.5	0~1	0~1	0~1	0~1	0.14~1	0.09~1	0.41~1	0.57~1	0.66~1	0.85~1
0.75	0~1	0~1	0~1	0~1	0~1	0~1	0.11~1	0.31~1	0.43~1	0.72~1
1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.06~1	0.20~1	0.58~1
2	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
Δ=0.5						**Δ=0.9**				
0.25	0.76~1	0.87~1	0.91~1	0.93~1	0.97~1	0.92~1	0.96~1	0.97~1	0.98~1	0.99~1
0.5	0.54~1	0.74~1	0.82~1	0.86~1	0.94~1	0.83~1	0.91~1	0.94~1	0.95~1	0.98~1
0.75	0.27~1	0.55~1	0.68~1	0.75~1	0.89~1	0.70~1	0.84~1	0.89~1	0.92~1	0.97~1
1	0.02~1	0.35~1	0.52~1	0.61~1	0.82~1	0.54~1	0.74~1	0.82~1	0.86~1	0.94~1
2	0~1	0~1	0~1	0.05~1	0.47~1	0~1	0.29~1	0.46~1	0.57~1	0.80~1
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.10~1	0.24~1	0.60~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.39~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.19~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.01~1

C_γ =1.0, w=0.1										
	Δ=0.1					Δ=0.3				
$n\rightarrow$ $C_x\downarrow$	5	10	15	20	50	5	10	15	20	50
w=0.25										
	Δ=0.1					Δ=0.3				
0.25	0~1	0~1	0~1	0~1	0.21~1	0.16~1	0.47~1	0.61~1	0.69~1	0.86~1
0.5	0~1	0~1	0~1	0~1	0~1	0~1	0.23~1	0.41~1	0.52~1	0.78~1
0.75	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.17~1	0.31~1	0.65~1
1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.09	0.50~1
2	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1

C_γ =1.0, w=0.25										
	Δ=0.5					Δ=0.9				
$n\rightarrow$ $C_x\downarrow$	5	10	15	20	50	5	10	15	20	50
0.25	0.58~1	0.77~1	0.84~1	0.88~1	0.95~1	0.85~1	0.92~1	0.95~1	0.96~1	0.98~1
0.5	0.38~1	0.63~1	0.74~1	0.80~1	0.91~1	0.76~1	0.87~1	0.91~1	0.93~1	0.97~1
0.75	0.13~1	0.45~1	0.59~1	0.68~1	0.86~1	0.62~1	0.79~1	0.85~1	0.89~1	0.95~1
1	0~1	0.25~1	0.43~1	0.54~1	0.79~1	0.46~1	0.69~1	0.78~1	0.83~1	0.93~1

C_γ =1.0, w=0.25										
Δ=0.5						**Δ=0.9**				
$n\rightarrow$ $C_x\downarrow$	**5**	**10**	**15**	**20**	**50**	**5**	**10**	**15**	**20**	**50**
2	0~1	0~1	0~1	0~1	0.41~1	0~1	0.21~1	0.40~1	0.51~1	0.77~1
3	0~1	0~1	0~1	0~1	0.04	0~1	0~1	0.03~1	0.17~1	0.56~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.33~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.13~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
***w*=0.5**										
Δ=0.1						**Δ=0.3**				
0.25	0~1	0~1	0~1	0~1	0~1	0~1	0.21~1	0.39~1	0.50~1	0.77~1
0.5	0~1	0~1	0~1	0~1	0~1	0~1	0.01~1	0.21~1	0.35~1	0.67~1
0.75	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.15~1	0.54~1
1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.38~1
2	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
Δ=0.5						**Δ=0.9**				
0.25	0.36~1	0.62~1	0.73~1	0.79~1	0.91~1	0.74~1	0.86~1	0.91~1	0.93~1	0.97~1
0.5	0.18~1	0.48~1	0.62~1	0.70~1	0.87~1	0.64~1	0.81~1	0.87~1	0.90~1	0.96~1
0.75	0~1	0.30~1	0.47~1	0.58~1	0.81~1	0.50~1	0.72~1	0.80~1	0.85~1	0.94~1
1	0~1	0.11~1	0.30~1	0.43~1	0.72~1	0.34~1	0.60~1	0.72~1	0.78~1	0.91~1
2	0~1	0~1	0~1	0~1	0.32~1	0~1	0.11~1	0.30~1	0.43~1	0.72~1

C_γ =1.0, w=0.25										
Δ=0.5						**Δ=0.9**				
$n\rightarrow$ $C_x\downarrow$	**5**	**10**	**15**	**20**	**50**	**5**	**10**	**15**	**20**	**50**
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.07~1	0.49~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.25~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.03~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
***w*=0.9**										
Δ=0.1						**Δ=0.3**				
0.25	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.14~1	0.28~1	0.63~1
0.5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.13	0.53~1
0.75	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.39~1
1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.23~1
2	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1

C_γ =1.0, w=0.9										
	Δ=0.5					Δ=0.9				
$n\rightarrow$ $C_x\downarrow$	5	10	15	20	50	5	10	15	20	50
0.25	0.10~1	0.42~1	0.57~1	0.66~1	0.85~1	0.60~1	0.78~1	0.85~1	0.88~1	0.95~1
0.5	0~1	0.29~1	0.46~1	0.57~1	0.80~1	0.49~1	0.71~1	0.80~1	0.84~1	0.93~1
0.75	0~1	0.12~1	0.31~1	0.43~1	0.73~1	0.35~1	0.61~1	0.72~1	0.78~1	0.91~1
1	0~1	0~1	0.15~1	0.28~1	0.63~1	0.18~1	0.49~1	0.63~1	0.71~1	0.87~1
2	0~1	0~1	0~1	0~1	0.19~1	0~1	0~1	0.18~1	0.31~1	0.65~1
3	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.38~1
4	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0.13~1
5	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1
6	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1	0~1

REFERENCES

EICHHORN, B.H. AND HAYRE, L.S. (1983), Scrambled randomization response methods for obtaining sensitive quantitative data, Journal of Statistical Planning and Inference, 7,307—316.

GUPTA, S., GUPTA, B. AND SINGH, S. (2002), Estimation of sensitivity level of personal interview survey questions, Journal of Statistical Planning and Inference, 100, 239—247.

SEARLS, D.T. (1964), The utilization of a known coefficient of variation in the estimation procedure, Journal of American Statistical Association, 59,1225—1226.

SEARLS, D.T. and INTARAPANICH (1990), A note on an estimator for the variance that utilizes the kurtosis, The American Statistician, 44(4), 295—296.

SINGH, S., JOARDER, A.H. and KING, M.L. (1996), Regrassion analysis using scrambled responses. Australian Journal of Statistics, 38(2), 201—211.

SINGH, H.P. AND MATHUR, N. (2003), Estimation of population mean when coefficient of variation of known using scrambled response technique. To appear in Journal of Statistical Planning and Inference.

WARNER, S.L. (1965), Randomized response: A survey technique for eliminating evasive answer bias. Journal of Statistical Planning and Inference, 60, 63—69.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1095—1103

AN OPTIMUM SIZE OF SAMPLING UNITS FOR TAJIKISTAN HOUSEHOLD SURVEY

Shyam Upadhyaya[1]

ABSTRACT

Determining the sample size is a crucial point in sampling design of a household survey. However, in the most commonly used two-stage cluster sampling design, we also need to decide the size (cluster) and number (PSU) of sampling units. For a given intra-cluster correlation coefficient, design effect is higher for larger cluster size, however, number of PSU's to be sampled is smaller, so the cost of the survey. On the other hand, smaller cluster size may produce more precise results, but with the higher survey cost. This article presents an attempt to find an optimum size of sampling units for the household survey in Tajikistan. The cost analysis showed that the number of PSU's could be significantly increased from the same budget if the necessary information on required parameters were known in advance.

Kew words: Two-stage design, design effect, intra-cluster correlation, effective sample size, cost ratio, optimum size.

1. Introduction

Often a household survey involves two-stage design with the enumeration area block as a primary sampling unit (PSU) selected from an area frame and the household as a secondary sampling unit (SSU) selected from the list. In such sampling design, PSU serves as a cluster and the household as its element. There may be different sampling schemes for selection of clusters and its elements. Again the most common design applies the PPS method at the first stage and a fixed number of samples at the second stage. Such scheme results in the self-weighting design as given below.

[1] Director, International statistical consultant (INSTAT), Kathmandu, Nepal; instat@wlink.com.np

The selection equation for the first stage is $f_1 = a\frac{m_j}{\sum m_j}$ and for the second stage $f_2 = \frac{b}{m_j}$. The selection probability at the first stage varies directly proportional to the size of PSU. When the fixed number of sample is selected at the second stage, selection probability varies inversely proportional to the size of PSU. Thus the overall rate of selection is given by:

$$f_0 = f_1 \times f_2 = \frac{ab}{\sum m_j} \to const. \text{ and } n = ab \quad (1)$$

Where, $\sum m_j$ – number of households in all PSU's (target population),

m_j – number of households in selected PSU and

b – number of households in sample per PSU (cluster size).

Despite its simplicity in implementation, there are some difficulties of attaining the self-weighting status of this design. First, household data in the frame may refer to the earlier period, which are used to select the PSU's, whereas the selection of household is made from the listing data collected just before the survey. Thus, m_j used in selection equation f_1 is not the same m_j that is in f_2 and cannot be mutually cancelled. The design may also not remain self-weighting throughout the estimation procedure due to the non-coverage and non-response. These problems are tackled in different ways, which are not covered in this article. Here, we discuss problems related to the allocation of samples over clusters.

2. Effect of cluster size on variance

With a predetermined sample size *n* and the desired constant number of *b,* the number of PSU's to be selected depends on the cluster size as $a = \frac{n}{b}$. Larger the cluster size smaller the number of PSU's in a sample and vice versa. The cluster size cannot be decided arbitrarily, because the design effect that inflates the variance in cluster sampling increases with the cluster size.

The population mean of *y* variable for the above type of self-weighting design can be estimated from the unweighted sample mean as given by:

$$\bar{y} = \frac{\sum_{\alpha=1}^{a}\sum_{\beta=1}^{b} y_{\alpha\beta}}{ab} \quad (2)$$

and its variance is estimated as;

$$Var\,(\bar{y}) = \frac{S^2}{n}[1 + (\bar{b} - 1)\rho] \tag{3}$$

where $\frac{S^2}{n}$ is the SRS variance and $[1 + (\bar{b} - 1)\rho]$ is the design effect (*deff*) which measures the effect of clustering the sample within PSU. Households clustered to a locality are more alike in terms of different socio-economic characteristics than those at large. The rate of homogeneity (*roh*) of the variable of interest within a cluster is measured by intra-class correlation coefficient denoted by ρ. The concept of deff and roh serves as the basis of discussion for an optimum size of PSU in sampling design.

As it can be seen, the design effect inflates the variance and its value depends on the cluster size and *roh*. Design effect of an estimated mean of *y* variable equals:

$$deff(\bar{y}) = 1 + (\bar{b} - 1)\rho \tag{4}$$

Normally *roh* tends to decline when increases the size of PSU, however significantly large size of PSU does not serve the purpose of clustering. In practice, a cluster may be a village, settlement or urban ward that have more or less clear boundaries to serve as a PSU. In such case, *roh* is present for the most of the important socio-economic variables. For a given value of *roh* design effect increases with the cluster size as shown below.

Table 1. Value of design effect for various cluster sizes (for $\rho = 0.05$)

b	5	10	15	20	25	30
deff	1.2	1.5	1.7	2.0	2.2	2.5

Thus smaller the cluster size lower the value of design effect. Design effect has the direct impact on the effective sample size. An ***effective sample size*** *(n_{eff})* is the number of observations required to produce the estimate with the same precision as that from a SRS design and computed as the ratio between the actual number of observations (survey sample size) *n* and the design effect – *deff.*

$$n_{eff} = \frac{n}{deff} \tag{5}$$

Effective sample size decreases as increases the design effect. A sample of 1000 households in a complex sample design with the value of *deff* = 2 can produce the estimates at the same level of precision, which would have been yielded from 500 households in a SRS design. Thus, in this case effective sample size is 500 households. Let us suppose, we want to produce the estimates from an effective sample size of 1000 households, how the survey sample size changes for different value of design effects. Obviously, there is no difference between former

and latter, when the design effect equals to 1, which mean that the complexity of design has no effect on variance.

Table 2. Required sample size for different value of *deff* (for $n_{eff} = 1000$)

deff	1	1.2	1.5	1.7	2.0	2.2	2.5
n	1000	1200	1500	1700	2000	2200	2500

By comparing the figures of Table 1 and 2, let us consider two sample plans for the same effective sample size (for $n_{eff} = 1000$). These plans are different in the number of PSU's, cluster size as well as overall sample size:

Sample plan I: $n_{(1)} = a_{(1)} b_{(1)} = 150 \times 10 = 1500$

Sample Plan II: $n_{(2)} = a_{(1)} b_{(1)} = 100 \times 20 = 2000$

The first plan has the sample size much smaller than the second plan, but it can produce the results effectively with the same precision as from the second plan. The gain is clearly achieved thanks to the smaller cluster size (lower design effect) of the first plan. However, it has greater number of PSU's, which may increase the survey cost. Decision in such situation very much depends on the per-unit cost of primary and secondary sampling units. For example, if the cost ratio of surveying PSU and SSU is higher than 10, sample plan II becomes cheaper than sample plan I despite the greater sample size in the latter one. Therefore, we can make an appropriate decision only after the cost analysis of the survey. How to find an optimum cluster size thereby determine the number of PSU's to be sampled? This issue is discussed in the example of a household survey conducted in Tajikistan.

3. Cost analysis of Tajikistan household survey

In recent years, the State Statistical Committee (SSC) of the Republic of Tajikistan conducted a number of national surveys. Although most of the surveys followed two-stage design, the cluster size varied across the surveys. Due to the lack of properly compiled cost information and estimates of variance, *deff* and *roh,* it was not possible to suggest an optimum cluster size thereby number of PSU's to be sampled in advance. The current exercise was done upon completion of the Poverty Reduction Monitoring Survey (PRMS) in 2003. So the main purpose of the analysis was to improve the sample design of similar household surveys in future. The cost analysis was based on a simple linear cost model for a two-stage sampling design given by:

$$C = C_0 + C_1 a + C_2 ab \tag{6}$$

Where,

a – number of PSU's in sample and

b – number of households selected per-psu (cluster size). For a sample design as given in formula (1) b is constant across all PSU's.
C_0 – Fixed cost
C_1 – Average cost of adding one PSU (a) in sample
C_2 – Average cost of adding one SSU (b) in sample

The total cost (C) presented in above model can be divided into two components: fixed cost and varying cost (field cost). Further analysis on the cluster size can be made by excluding the fixed cost. As we are interested actually in the cost ratio, effect of the omission of fixed cost, which is almost equal for both of C_1 and C_2 can be ignored. Then the cost of data collection C_f is given as,

$$C_f = C_1 a + C_2 ab \tag{7}$$

The cost analysis of the household survey of this kind in Tajikistan was done for the first time. PRMS was conducted under a technical assistance project of Asian Development Bank. In the cost analysis, fees of international consultants as well as contribution of the ADB staff members in drafting the questionnaire and producing analytical tables were excluded. There were some difficulties in distribution of cost items to the fixed and varying costs of surveying PSU and SSU. For example, same staffs were involved in designing the survey that included sampling design, preparing questionnaire and instruction manual, tabulation plan, field staff training etc. It was difficult to split their time subsequently the cost to different survey activities. Therefore, salaries of the central staffs were taken as the cost of survey design. Under the training cost, travel expenses of supervisors and enumerators to the capital city were taken as the cost of PSU and households respectively. Survey materials and stationery were considered as fixed cost although the cost of printing of the questionnaire could vary depending on the sample size. The main cost varying to the number of PSU's and households are the allowances paid to enumerators and supervisors. Supervisors had overall responsibility for the whole PSU including the selection of households from the list, therefore their number depended on the number of PSU. Allowances were paid on daily basis without specifying the proportion of salaries and per-diems. Details of cost classification are given below.

Table 3. Cost structure of PRMS by components[1]

Components	Fixed cost	Varying cost depending on sample size	
		Cost of surveying PSU	Cost of surveying households
A. Survey design (Salary of the central staffs)	XXXX		
B. Printing cost (questionnaire, manuals etc)	XXX		
C. Training cost: Cost of training materials and others Travel and daily allowance paid to supervisors Travel and daily allowance paid to enumerators	XXX	XXX	XXXX
D. Field operation cost Central supervision (cost of travel of central staffs) Allowances of field supervisors Allowances of interviewers Transportation cost to PSU Transportation within PSU Survey materials (bags, files, stationery)	XXX XXX	XXXX XXX	XXXXX XXXX
E. Data processing cost	XXXX		
F. Publishing and dissemination	XXXX		
Total cost	C_0=XXXXX	C_{PSU} = XXXXX	C_{HH} = XXXXX

Having the cost data presented in the format of the Table 3, following indicators were computed for the cost model of type (6).

[1] It has not been possible to present the detail figures in this table, however cells are marked to show the distribution. The cost ratio and other indicators are presented in actual figures.

Table 4. Estimated value of major cost indicators

Indicators	Computation	Value in US$
Average cost of adding one PSU	$C_1 = \frac{C_{PSU}}{a}$	39.8
Average cost of adding one household	$C_2 = \frac{C_{HH}}{n}$	7.4
Average survey cost per PSU	$C_{psu} = C_1 + bC_2$	187.8
Average survey cost per interview (b=20)	$\overline{C}_f = C_1 / b + C_2$	9.4

The cost ratio between PSU and SSU C_1 / C_2 was 5.4. The low cost ratio is explained by the level of remuneration in general in the country.

4. Optimum cluster size

An optimum cluster size is the one that minimizes the variance (formula 3) and the cost per interview given above. Both of these components are expressed as:

$$Var\overline{C}_f = V(1+(b-1)\rho)(C_1 / b + C_2), \text{ where } V = \frac{S^2}{n} \qquad (8)$$

From the above relation L. Kish (1965) derived the formula for an optimum size b_{opt}:

$$b_{opt} = \sqrt{\frac{C_1}{C_2}\frac{(1-\rho)}{\rho}} \qquad (9)$$

For the given cost ratio the optimal cluster size changes for different values of *roh*, so changes the minimised value of $Var\overline{C}_f$. The following table presents the minimised value of $Var\overline{C}_f$ for $b = b_{opt}$ (optimum size) and $b = 20$ (actual size). Relative efficiency of design is calculated as the relation of $Var\overline{C}_f$ for optimised size to the actual cluster size. For comparison, $Var\overline{C}_f$ is also presented for $b = 10$. Relative efficiency of our design with the cluster size of 20 households is higher only for the variables with *roh* values up to 0.05. However, most of the estimates had *roh* values ranging from 0.05 to 0.1. So, the relative efficiency would have been much better for most of the estimates if the cluster size were 10 households.

Table 5. Optimal cluster sizes for different values of *roh* and relative efficiency of the design

	Value of *roh*					
	0.01	0.05	0.075	0.1	0.15	0.2
b_{opt}	23	10	8	7	6	5
$Var\overline{C}_f \ (b = b_{opt})$	*V* 11.14	*V* 16.50	*V* 18.87	*V* 20.94	*V* 24.52	*V* 27.61
$Var\overline{C}_f \ (b = 20)$	*V* 11.17	*V* 18.31	*V* 22.77	*V* 27.23	*V* 36.15	*V* 45.07
$Var\overline{C}_f \ (b = 10)$	*V* 12.40	*V* 16.50	*V* 19.06	*V* 21.62	*V* 26.74	*V* 31.86
Relative efficiency in % for (b=20)	99.69	90.12	82.87	76.89	67.81	61.26
Relative efficiency in % for (b=10)	89.80	100.00	99.00	96.83	91.67	86.65

When the cluster size is reduced, number of PSU to be sampled for given sample size is increased. How many PSU could be surveyed from the given budget? Let us replace *b* of formula (7) by b_{opt} and calculate the total cost of survey of each PSU as: $C_{psu} = C_1 + b_{opt} C_2$. Now, we divide the survey budget by the cost of each PSU to get the number of PSU that can be sampled for the given amount.

$$a_c = \frac{C_f}{C_1 + b_{opt} C_2} \tag{10}$$

After some adjustments and rounding up it was found that for similar household survey of the same sample size and budget a sampling plan of $\{a = 500 \quad b = 12 \quad n = ab = 6000\}$ would be much preferable to the previous sampling design with $\{a = 300 \quad b = 20 \quad n = ab = 6000\}$. Smaller number of PSU as a cost cutting measure was not fully justified, however it was not obvious at the design stage as the information on required parameters was not available.

Above discussions on optimum size mainly touch upon the general questions. The design could be further improved if the cost of data collection C_f is differentiated to urban-rural domains. In a mountainous country like Tajikistan, there are remote areas not easily accessible by public transport. The travel cost between sample points in such areas is often higher than those in the urban PSU's. It might be cost effective to make cluster size larger in rural and smaller in urban areas. Smaller cluster size is preferable in urban area also because in many developing countries cities are often segregated to slum and posh areas. In such conditions major variables observed in socio-economic surveys are most likely to be correlated within the clusters. Additionally, sample allocation also depends on the availability of manpower, logistic supports and capacity of field offices, which need due consideration in survey organisation.

5. Conclusion

While planning the survey we often face a difficult choice between the precision of expected results and the cost. What would be an optimal solution? Household surveys not only produce the statistics on study variables as per demand of users which are collected by administering the questionnaire, but also generate a lot of important information that can be used in future in finding a survey design that has a good balance of the precision and cost. However, in many developing countries the latter part of information is not properly analysed or documented. An optimal size of sampling units can be determined a priori, if the necessary information on cost as well as on variance, intra-cluster correlation and design effect at least for key variables are available from earlier surveys.

Perhaps the best way of preserving such information is to estimate the variance and cost related statistics immediately after the survey. In Tajikistan, these statistics were presented in the Technical report of PRMS, which was produced separately from the analytical report.

REFERENCES

COCHRAN W. (1977) *Sampling Techniques*, John Wiley and Sons, New York.

KISH L. (1965) *Survey Sampling*, John Wiley and Sons, New York.

PETTERSSON H. and BOUNTHAVY S. (2002) A cost model for an income and expenditure survey, *http://unstats.un.org/unsd/HHsurveys/ch13r2_web.pdf*

STATE STATISTICAL COMMITTEE OF TAJIKISTAN, (2003), The Technical Report of the Poverty Reduction Monitoring Survey, Dushanbe.

UPADHYAYA S. (2003) Creating survey blocks from household register, Proceedings of 54th ISI Session, Berlin.VERMA V (1990) *Sampling methods*, SIAP, Tokyo.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1105—1109

A NEW APPROACH TO πPS SAMPLING SCHEME-I

Manju Saxena and Alpana Saxena[1]

ABSTRACT

In the paper a sampling scheme using auxiliary variable for cluster sampling has been proposed. The auxiliary variable is used in selecting the elements in clusters. The efficiency of the proposed sampling scheme is compared to the usual cluster sampling with equal probabilities of selection as well as to SRSWOR. It has been observed that of all the schemes proposed sampling scheme which make use of πPS approach is the best.

Key words: Horvitz-Thompson estimator, Midzuno-Sen sampling scheme, auxiliary variable, inclusion probability, cluster sampling, efficiency.

1. Introduction

Cluster sampling is a well-known sampling scheme. It is adopted not only where elements in the population are unidentifiable but also due to other reasons such as reduction in cost, simplicity etc. There are various ways of forming clusters. In this paper we propose a sampling scheme, which make use of an auxiliary variable in selecting the elements in the clusters. The usual sampling scheme makes use of no such information. The inclusion probability for individual and pairwise units have been worked out. The efficiency of this scheme has been investigated empirically.

2. Proposed Sampling Scheme

Let the population consists of M distinct and identifiable units and N clusters with M_t units in the t^{th}- cluster.

(1) Select n clusters from N clusters by SRSWOR.
(2) From each of the selected cluster select m units by Midzuno (1951) or Sen (1953) sampling scheme.

[1] Department of Mathematics and Statistics, J.N.V. University, Jodhpur, 342 005, India.
e-Mail: manjusairam@rediffmail.com

The proposed sampling scheme makes use of an auxiliary variable in selecting the elements in the clusters.

3. Calculation of Inclusion Probabilities

The inclusion probability of a unit can simply be obtained by first multiplying the probability of selection of cluster to which the desired unit belongs at step (1) by probability of selecting the desired unit from the cluster at step (2) considering the repetitions also by adding over clusters.

The inclusion probability of i^{th} unit belonging to t^{th} group is given by

$$\pi_i^t = \frac{1}{N_{C_n}} \left[\sum_{\substack{t \\ (i \in t)}}^{N} I_{ti} - \sum_{\substack{t,t' \\ (i \in t,t')}}^{{}^{N}C_2} I_{ti} I_{t'i} - \sum_{\substack{t,t',t'' \\ (i \in t,t',t'')}}^{{}^{N}C_3} I_{ti} I_{t'i} I_{t''i} \right] \tag{1}$$

Where

I_{ti} – Inclusion probability at step (2) of i^{th} unit belonging to t^{th} cluster equal to

$$\left[\frac{(M_t - m)}{(M_t - 1)} p_{ti} + \frac{(m-1)}{(M_t - 1)} \right]$$

and

$$p_{ti} = \frac{X_{ti}}{\sum_{p=1}^{Mt} X_{tp}} = \frac{X_{ti}}{X_t}$$

X_{ti} is the value of the auxiliary variable corresponding to i^{th} unit belonging to the t^{th} cluster, t = 1, 2, …, N and = 0 otherwise,

X_t is the sum of the auxiliary variable belonging to t^{th} cluster.

The inclusion probability of two units belonging to the same cluster can be obtained by adding over clusters the product of the probability of selecting the desired cluster at step (1) to the probability of selecting the desired pair of units at step (2).

$$\pi_{ij}^{tt} = \sum_{\substack{t \\ (i,j \in t)}} \frac{1}{N_{c_n}} \left[\frac{(m-1)(m-2)}{(M_t - 1)(M_t - 2)} + \frac{(m-1)(M_t - m)}{(M_t - 1)(M_t - 2)} (p_{ti} + p_{tj}) \right] \tag{2}$$

Similarly the inclusion probability of two units belonging to two different clusters can be obtained by multiplying the probability of selection of desired pair of clusters at step (1) by the probability of selecting one unit each from the two selected clusters at step (2).

$$\pi_{ij}^{tt'} = \frac{n(n-1)}{N(N-1)}\left[\frac{(M_t - m)}{(M_t - 1)}\, p_{ti} + \frac{(m-1)}{(M_t - 1)}\right] \mathrm{x} \left[\frac{(\mathrm{M}_{t'} - m)}{(M_{t'} - 1)}\, p_{t'j} + \frac{(m-1)}{(M_{t'} - 1)}\right] - A.F. \qquad (3)$$

where A.F. is the adjustment factor which is the contribution from within cluster component if any.

A.F. – Probability of selecting the required pair (i,j) of units from within cluster (say t) component to the probability of selecting one of these clusters from between cluster component.

$$\sum{}' \frac{1}{N_{C_n}}\left[\frac{(m-1)(m-2)}{(M_t - 1)(M_t - 2)} + \frac{(m-1)(M_t - m)(p_{ti} + p_{tj})}{(M_t - 1)(M_t - 2)}\right] \mathrm{x} \left[\frac{(M_t' - m)}{(M_t' - 1)}\, p_{t'i} + \frac{(m-1)}{(M_t' - 1)}\right] \qquad (4)$$

Σ' is taken over all such terms.

4. Estimation Procedure

After obtaining π_i and π_{ij} we can use the Horvitz Thompson estimator for estimating population total $Y = \sum_{i=1}^{N} Y_i$. The Horvitz Thompson estimator of population total is given by

$$\overset{\Lambda}{Y}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i} \qquad (5)$$

The variance of the Hortvitz Thompson estimator is given by

$$V\left(\overset{\Lambda}{Y}_{HT}\right) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i}\, Y_i^2 + \sum_{i \neq j}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i\, Y_j \qquad (6)$$

The estimate of the variance of the Hortvitz Thompson estimator is given by

$$\overset{\Lambda}{V}\left(\overset{\Lambda}{Y}_{HT}\right) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i}\, y_i^2 + \sum_{i \neq j}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i\, y_j \qquad (7)$$

The usual cluster sampling consists of the following steps:

(i) Select ‘n’ clusters out of ‘N’ clusters by SRSWOR.

(ii) From each of the ‘n’ selected clusters select ‘m’ units by SRSWOR.

Substituting $P_{ti} = 1/M_t$ in expressions (1), (2), (3) and (4) the inclusion probabilities for individual and pairwise units under usual cluster sampling can be obtained.

It is difficult to compare theoretically the efficiency of the proposed cluster sampling scheme with usual cluster sampling scheme and SRSWOR. However an exhaustive empirical comparison can be attempted.

For comparison we have considered three populations which are similar to those considered by Cochran (1977). For the first population Y/X increases as x increases, for the second Y/X decreases as X increases whereas for the third population Y/X and X have no correlation. The used populations are given in the following table.

Table 1.

Population	**(a)**		**(b)**		**(c)**	
S.No.	**X**	**Y**	**X**	**Y**	**X**	**Y**
1	2	16	2	28	2	16
2	2	17	2	27	2	17
3	3	27	3	39	3	27
4	3	28.5	3	37.5	3	28.5
5	4	40	4	48	4	40
6	4	42	4	46	4	42
7	5	55	5	55	5	55
8	6	69	6	63	6	63
9	6	72	6	60	6	60
10	7	87.5	7	66.5	7	66.5
11	7	91	7	63	7	63
12	8	108	8	68	8	68
13	8	112	8	64	8	64

These populations have been rearranged into 4 clusters (overlapping) of 4 units each as shown in the following table.

Table 2.

Cluster Number	**Units in the Cluster**
1	1, 2, 3, 4
2	4, 5, 6, 7
3	7, 8, 9, 10
4	10, 11, 12, 13

Variances of the proposed sampling scheme, usual sampling scheme and also for SRSWOR have been worked out for n = 2 and m = 2. The variances are presented in the following table.

Table 3. Variances of sampling schemes considered

Population	$V_{(SRSWOR)}$	V(usual)	V(proposed)
(a)	0.7546	0.7497	0.7337
(b)	0.1551	0.1917	0.1863
(c)	0.2521	0.2573	0.2509

These results indicate that the performance of the proposed sampling scheme is highly satisfactory and the gain in efficiency over the existing scheme is quite considerable. Taking into account various factors such as cost, simplicity, efficiency etc. the proposed sampling scheme can be used in practice with advantage.

REFERENCES

COCHRAN W. G. (1977): *Sampling Techniques*, Third edition, John Wiley and Sons, Inc., New York, London.

GOEL and Singh (1977): On formulation of clusters, *Journal Indian Soc. Agri. Stat*. Vol. 1.

HORVITZ, D. G. and THOMPSON, D. J. (1952): A generalization of sampling without replacement from a finite universe, *Journal Amer. Stat. Assoc*. Vol. 47, pp. 663—685.

MIDZUNO, H. (1951): "On the sampling system with probability proportionate to sum of sizes". *Ann, Inst. Stat. Math*. 2, pp. 99—108.

SAXENA, M. (1984): "Sampling with varying probabilities without replacement Ph.D. Thesis, University of Jodhpur.

SEN, A.R. (1953): "On the estimate of variance in sampling with varying probabilities", *Journal. Ind. Soc. Agri. Stat*., No. 5, pp. 119—127.

SRIVASTAVA, A.K. (1996-1997): "On overlapping clusters and Multiplicity sampling", *Journal Ind. Soc. Ag. Stat*. No. 49, pp. 307—318.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1111—1119

MANIPULATION OF THE IRRESOLUTE VOTERS IN POLITICAL SAMPLE SURVEYS: A MATHEMATICAL PROCESS

Nicolas Farmakis[1]

ABSTRACT

In sample surveys designed to search for voting intentions, the problem of the manipulation of the irresolute voters (voters who have not decided about their voting, "floating" voters) arises. A method for managing their distribution is proposed in this paper. The method is very simple and is based on certain empirical coefficients from the Greek political experience of the last 25 years.

***Key words*:** Sampling, survey, irresolute voters, voting.

1 Introduction

In sample surveys of voting intention we sometimes use the following procedure. A ballot box reaches a **random sample** of voters (for random sampling see in Cochran, 1977, and Farmakis, 1992, 2000) asking for their opinion. Every voter has to respond to the following two questions written on the ballot (possibly among some other questions of demographic interest, like age, sex, etc.):

1."For which party did you vote at the previous election?"
2."For which party are you going to vote at the next election?"

At the end we open the ballot box and we register the flow of the votes from the parties in the previous election to those in the next election. Some voters may write the word "irresolute" on the ballot as an answer to the second question (i.e. they declare "I have not decided yet") or may give some other answers like "white ballot", "no answer", etc. In the next paragraph, we will try to redistribute these

[1] Aristotle University of Thessaloniki, Department of Mathematics, GR-54124, Thessaloniki, Greece; email: farmakis@math.auth.gr

ambiguous answers to the actual parties in the next election by a simple Mathematical method of a proportional nature.

2 The method and an algorithm

Suppose we had *m* political parties, K_j, j=1, 2,…, *m*, in the previous election. The small parties are usually grouped together into the *m*-th 'party'. Suppose also that there will be *r* parties, k_i, *i*=1, 2, …, *r*, for the next election. Every voter in the sample has basically three choices, when answering the second question:

a) to vote for one of the r real parties;
b) to write the word "irresolute" (meaning "I have not decided yet"); or
c) to give any other answer like "white ballot", "no answer", etc.

Some of the voters are not really irresolute ones even if they declare "irresolute". They do it for personal reasons. Anyway all the irresolute voters, in the day of election, have to decide and become resolute ones. In accordance with our experience in Greece for the last 25 years, this is true for more than 75% of them. Also a little percentage (at most 20% according to our experience) of the voters giving the ambiguous answers described in the third case above, will vote for one of the r parties in the next election.

Let $100 \cdot f_j$, where $\sum_{j=1}^{m} f_j = 1$, $0 \le f_j \le 1$, j=1, 2, …, *m*, represent the percentages of votes gained by each of the *m* parties in the previous election. For every *j*=1, 2,…, *m* let $100 \cdot \bar{p}_{ij}$, *i*=1, 2, …, *r* be the percentage flow of $100 \cdot f_j$, from the j[th] party in the previous election to the i[th] party in the forthcoming election, $0 \le \bar{p}_{ij} \le 1$ (Table 2.1 at the end of this section). Since we have $\bar{p}_{r+1,j} \ge 0$ and $\bar{p}_{r+2,j} \ge 0$, it is in general $\sum_{i=1}^{r} \bar{p}_{ij} \le 1$, $\forall$ j=1, 2, …, *m*.

We now have to distribute (main idea for the present paper) the $100 \cdot \lambda_1$%, $\lambda_1 \in [0.75, 1.00]$, of the irresolute voters and the $100 \cdot \lambda_2$%, $\lambda_2 \in [0.0, 0.2]$, of those giving other ambiguous answers, to the real parties of the next election. This is because some of the irresolute voters are not really irresolute ones and some of them will select one of the actual r parties on election day. Thus, we obtain a new distribution for the flow rate that is closer to the reality, as follows:

$$\tilde{p}_{ij} = \hat{p}_{ij} \cdot \left\{ 1 + \frac{\sum_{q=1}^{2} \lambda_q \cdot \hat{p}_{r+q,j}}{\sum_{h=1}^{r} \hat{p}_{hj}} \right\}, \quad \text{i=1, 2, …, r and j= 1, 2, …, m,} \qquad (2.1)$$

The above flow rates will help us to make point estimations of the voting percentages in the next election according to the following ***algorithm:***

Step 1. Read f_j *and* $\hat{p}_{ij}$, *i=1, 2, ..., r+2, j=1, 2, ..., m (Table 2.1) and* λ_1, λ_2.

Step 2. Calculate $\tilde{p}_{ij}$, *i=1, 2, ... r, j=1, 2,..., m, according to (2.1), for suitable (given) values of* λ_q, *q=1,2.*

Step 3. Calculate $P_i = \sum_{j=1}^{m} f_j \cdot \tilde{p}_{ij}$, *i=1, 2,..., r and* $S = \sum_{i=1}^{r} P_i$.

Step 4. If S=1 (a very-very seldom case), then $\Phi_i = P_i$, *i=1, 2,..., r , else*

$$\Phi_i = \frac{P_i}{S}, \text{i} = 1,...,\text{r}$$

Step 5. Write 100 Φ_i %, i=1, 2, …, r, the point estimations for the vote percentages for the parties in the next election. End of the algorithm.

The key concept here is the two numbers λ_1, λ_2 with their values in well-known subintervals of $[0,1]$. The subintervals are well known to the researchers from their everyday experience as they work on political sampling surveys for many years. So they have checked their predictions against the final results of the suitable election too many times.

Remark 2.1. What about the new generation of voters?

If from one election to the next one there are reasonable new registered voters, e.g. more than a 5% of the electoral corps, we have to find the suitable percentages for them, as those figuring above in Step 5. Let us denote these percentages as

$$100 \cdot \varpi_i \% \quad \text{i=1,2,3,…,r}$$

and suppose that the weights of the old and the new registered voters in the electoral corps are w_1 and w_2 respectively, with $w_1+w_2=1$ and $w_1, w_2 \in [0,1]$. Then the final announcement of the percentages (Step 5 of the algorithm) will be the next:

$$100 \cdot \varphi_i \% = 100 \cdot (w_1 \cdot \Phi_i + w_2 \cdot \varpi_i)\% \quad \text{i=1,2,3,…,r.} \qquad (2.2)$$

If the elections take place 4 years after the previous one, then the weight of the new registered voters in Greece is

$$f_{new}=w_2 \in [0.05,0.08].$$

We note with emphasis that the new registered voters are the most difficult part of the voting population for estimation and prediction. This is because:
1) The researchers face this subpopulation for the first time.
2) There is also the famous "gap of generations" to be in account.
We can sincerely and with no doubt say that:
"It is difficult to guess the new voters behavior and so we have to be careful".

Remark 2.2. The best choice seems to be the adoption of different values for the coefficients λ for every column in Table 2.1, i.e. 2m values for old registered voters and 2 values more for the new registered ones. This is difficult, even if it seems to be a more precious choice, because is always difficult to check the successfulness of the choice –facing the final results of the election- for too many parameters. We have to study more and more on it, yet.

The proposed method was applied in The General Elections of September 1996 in Greece in order to predict the election results in the second electoral area of Athens -called in Greek **"B' Athinon"** — with very good precision. Some results are shown as "Example 2" in the next section of the paper.

Table 2.1.

$f_j \rightarrow$ | f_1	f_2	f_3	. . .	f_m / $\sum_j f_j = 1.00$	
$k_i \backslash K_j$	K_1	K_2	K_3	. . . K_m	
k_1	$\bar{p}_{11}$	$\bar{p}_{12}$	$\bar{p}_{13}$	. . . $\bar{p}_{1m}$	
k_2	$\bar{p}_{21}$	$\bar{p}_{22}$	$\bar{p}_{23}$	. . . $\bar{p}_{2m}$	
k_3	$\bar{p}_{31}$	$\bar{p}_{32}$	$\bar{p}_{33}$	. . . $\bar{p}_{3m}$	
.	…	…..	. . .	…..	
.	…	…..	. . .	…..	
	…	…..	. . .	…..	
k_r	$\bar{p}_{r1}$	$\bar{p}_{r2}$	$\bar{p}_{r3}$	. . . $\bar{p}_{rm}$	
Irresolute	$\bar{p}_{r+1,1}$	$\bar{p}_{r+1,2}$	$\bar{p}_{r+1,3}$	. . . $\bar{p}_{r+1,m}$	
White, etc.	$\bar{p}_{r+2,1}$	$\bar{p}_{r+2,2}$	$\bar{p}_{r+2,3}$	. . . $\bar{p}_{r+2,m}$	
Total	1.00	1.00	1.00	. . . 1.00	

Table 2.1. shows the flow rates of voters from the m parties of the previous election to the r parties of the next election. It also shows the rates of irresolute

voters and of those giving ambiguous answers like: "No answer" or "White ballot", etc.

3 Two Examples

Example 1 (artificial data)

Suppose that in the previous election we had m=3 "big" political parties and a number of other parties which we will present as a 4th one under the title "others". For the soon coming election we have r=2 parties and as a 3rd one all the others under the title "others" too. A week before the election we make a sample survey with results shown in Table 3.1. The new registered voters are the 6% of the whole population. Try to give a point estimation of the results of the next election.

Table 3.1.

f_j	0.40	0.35	0.20	0.05	1.000	
i\j	1	2	3	4=m(others)	Total	***New registered***
1	150	18	12	4	184	10
2	15	160	13	2	190	11
3=r	5	7	82	25	119	6
Irresolute	50	50	10	5	115	9
White, etc.	30	40	8	4	82	4
Total	250	275	125	40	690	40

The total size of the sample is n=690+40=730 voters

Answer: First of all we choose the coefficients $\lambda_1=0.8 \in [0.75,1.00]$ for the irresolute voters and $\lambda_2=0.1 \in [0.0,0.2]$ for the other answers. After this we construct Table 3.2 with suitable rates $\bar{p}_{ij}$ instead of the absolute numbers of the table 3.1. The method is based on column distribution of voters, i.e. $\sum_{i=1}^{r+2} \bar{p}_{ij} = 1$, j=1,2,3,4. We do the same for the new registered voters (see last columns of Table 3.1 and Table 3.2). By the help of Table 3.2 we calculate the values $\tilde{p}_{ij}$, i=1,2,3 and j=1,2,3,4, in accordance with (2.1). We are based on values of λ's, indicated in table 3.2 in suitable cells of its 3rd column. We put all the above results in the Table 3.3. We also do the same for the new voters –using the same values for λ's- and we put the results in the suitable cells of the last column in Table 3.3.

Table 3.2.

f_i	0.40	0.35	0.20	0.05	1.000	
i\j	1	2	3	4=m		***New Registered***
1	0.600	0.065	0.096	0.100		0.250
2	0.060	0.582	0.104	0.050		0.275
3=r	0.020	0.025	0.656	0.625	Values for λ	0.150
Irresolute	0.200	0.182	0.080	0.125	λ_1=0.8	0.225
White, etc.	0.120	0.145	0.064	0.100	λ_2=0.1	0.100
Total	1.000	0.999	1.000	1.000		1.000

Flow rates from the m parties of the previous election to the parties of the forthcoming election. Distribution based on columns.

Table 3.3.

f_i	0.40	0.35	0.20	0.05		
i\j	1	2	3	4=m	$\sum_j f_j \cdot \tilde{p}_{ij}$	***New Registered***
1	0.752	0.080	0.104	0.114	0.355	0.320
2	0.075	0.721	0.113	0.057	0.308	0.352
3=r	0.025	0.031	0.710	0.714	0.199	0.192
Total	0.852	0.832	0.927	0.885	0.862	0.864

The last two columns of Table 3.3 are going to give the estimations for old and new voters respectively.

From the last two columns of Table 3.3 we calculate -according to Step 4 of the algorithm- the values of estimations Φ_i for the old voters and the respective values ϖ_i for the new registered ones, where i=1,2,3. The suitable weights are w_1=0.94 and w_2=0.06 and according to (2.2) we have the estimation of percentages $100 \cdot \varphi_i\% = 100 \cdot \left(0.94 \cdot \Phi_i + 0.06 \cdot \varpi_i\right)\%$, i=1,2,3 for the three parties of the next week election. All these results are shown in Table 3.4.

Table 3.4.

Weights →	w_1=0.94	w_2=0.06	
i ↓	Φ_i	ϖ_i	$\varphi_i = w_1 \cdot \Phi_i + w_2 \cdot \varpi_i$
1	0.412	0.370	0.4095 or 40.95%
2	0.357	0.407	0.3600 or 36%
3=r	0.231	0.222	0.2305 or 23.05%
Total	1.000	0.999	1.0000 or 100%

We usually announce the results (percentages) in the 0 to 100 scale as it is shown in the second part of the last column of Table 3.4

Example 2 (real research data)

In the present example we refer to a real search which took place early in September 1996, in order to predict the results for "The General Elections '96", in Greece. Since we know now the final results we can check the reliability of the above method. We refer only to electoral area named in Greek "B' Athinon". The political parties, which took part in the election were:

"PASOK", "NEA DIMOKRATIA" abr. "N.D.", "POLITIKH ANOIXI" abr. "POL.AN.", "KKE", "SYNASPISMOS" abr. "SYN", "DIMOKRATIKO KINONIKO KINIMA" abr. "DIKKI" and finally all the other "little" parties under the title "OTHERS". We had a lot of data collected in a search by telephone. Responsible Co.: "HELLENIC RESEARCH HOUSE". We reached the problem of estimating the percentages for the political parties using as a classifying parameter the vote in previous election, the sex, the age, etc. The best results were those ones based on the parameter "Vote" (in previous election) and the parameter "Sex". In the Table 3.5, we give a short message on the reliability of the method.

Table 3.5.

Political Parties	Estimations by "VOTE"	Estimations by "SEX"	***Real Final Results of Elections***
PASOK	0.3890	0.3830	**0.3877**
N.D.	0.3626	0.3286	**0.3206**
POL.AN.	0.0377	0.0377	**0.0265**
KKE	0.0521	0.0787	**0.0835**
SYN	0.0584	0.0764	**0.0867**
DHKKI	0.0595	0.0602	**0.0589**
OTHERS	0.0407	0.0356	**0.0361**
Total	**1.0000**	**1.0000**	**1.0000**

Estimating the results of General Elections'96 in Greece by using "Vote" as a classifying parameter.

4. Discussion

The above method seems to be better, if we separate the voters on the basis of sex instead of the previous voting preference (behavior). If we work on the basis of the sex we have nothing to do with the new generation behavior, of course. The duty to decide about the weights of the old and new voters is always a heavy one.

It is useful to note here that the above method, in a more general scheme, can be consisted of the three basic next steps:

(A) Separate the Electoral Corps in m groups via some property of its persons, sex, previous voting behavior, age, economic status, etc.

(B) Make a Registration of the flow rates from the groups to the political parties of the next elections and to the several ambiguous answers.

(C) Follow the algorithm described in section 2.

The problem of definition of the value for the coefficient w_1 and w_2 in (2.2) is open for many cases, due to generation's gap in behaviour. For Greece we have some good enough estimation, like this one in Remark 2.1, and we try to be better on this. Every country is different of course. In any case if we use as w_1 and w_2 the population percentages of the old voters and of the new ones we are in a good way, supposing that all the voters will go to vote or the two strata will have the same percentage of missing votes.

A very interesting idea is to have some kind of space estimations. There are basically two kinds of spaces estimating the percentages gained by political parties:

The 1st of them is the classic one confidence interval (C.I.)

$$100\cdot(1-\alpha)\%\ \text{C.I. } of\ \ \varphi_k=\{\varphi_k \pm SEE\},\ \text{k=1, 2, 3, …,r} \tag{4.1}$$

where SEE= $z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\varphi\cdot(1-\varphi)}{n-1}}$, as in Cochran (1977, Chapter 3), is described.

The 2nd one is the one obtained by a computer running. Computer runs the algorithm of section 2 of this paper for a very dense net of values of $\lambda_1 \in [0.75,1.00]$ and of $\lambda_2 \in [0.00,0.20]$, e.g. for a step of size h=0.0025. We order the computer to give as an output the max and min values obtained for all the percentages φ_k, k=1,2,3,…,r. The space $[\varphi_{min}, \varphi_{max}]$ could be a kind of confidence interval of the percentage φ_k of the suitable party. This interval is an empirical one and it has not the properties of a real confidence interval. Anyway after some years of successful experience we trust on it.

Acknowledgment

We thank "HELLENIC RESEARCH HOUSE Co" for giving to the author the data of the research, in order to be used for a checking (for one more time) of the reliability of the method proposed here.

REFERENCES

COCHRAN, W. (1977) *Sampling Techniques* New York: John Wiley & Sons.

FARMAKIS, N. (1992 & 2000) *Introduction to Sampling.* Thessaloniki: A&P Christodoulidis, (in Greek).

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1121—1129

STOCHASTIC ORDERS AND CLASSES OF PRIOR DISTRIBUTIONS

Marek Męczarski[1]

ABSTRACT

The usual stochastic order and the dispersive order are considered in the set of all probability distributions as tools to construct classes of prior distributions. These classes are widely exploited in Bayesian robust statistical methods.

***Key words*:** usual stochastic order, dispersive order, classes of prior distributions, robust Bayesian statistical procedures.

1. Introduction

The problem of choice of a prior distribution in statistical Bayesian analysis is addressed by rich literature (see e. g. Bernardo and Smith, 1994). They are reflected by constructing classes of the prior distributions to be considered in a statistical model of interest. The classes of priors allow to perform robustness (stability) analyses for statistical decision rules (see Męczarski, 1998; Ríos Insua and Ruggeri, 2000). The construction of classes of priors is treated in many papers and several book chapters, because such a family should satisfy the following informal conditions:

1^o the principle of the construction should be intuitively clear and easy to interpret also for non-experts having some general knowledge on probability;

2^o computations for stability studies (extremum points and values of statistical functionals or their differential characteristics) should be easy to made, to solve explicitly and their results should be easy to interpret.

The general form of the statistical functional is as follows: let us consider a statistical model $(\mathcal{X}, \mathcal{M}, \mathcal{P})$, where $\mathcal{X}$ is a sample space, $\mathcal{M}$ — a σ-algebra of events in $\mathcal{X}$ and $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ is a family of probability distributions.

[1] Institute of Econometrics, Warsaw School of Economics, Al. Niepodległości 162, 02-554 Warszawa, Poland.

Moreover, we assume that there exists the probability space $(\Theta, \mathcal{F}, \Pi)$, where Π is a prior distribution. Then we define

$$S(\pi \mid x) = \frac{\int_{\Theta} g(x, \vartheta) f(x \mid \vartheta) \pi(\vartheta) d\vartheta}{\int_{\Theta} f(x \mid \vartheta) \pi(\vartheta) d\vartheta},$$

where g is a real function with the arguments from $\mathcal{X} \times \Theta$ corresponding to the decision problem of interest (e. g. for g being a loss function we obtain the posterior risk) and $f(x \mid \vartheta)$ denotes a sample density function.

To define classes of priors a structure is needed in the set of prior distributions. We will consider classes of priors such that they can be described in terms of stochastic orders.

2. Stochastic orders

Stochastic orders may be understood as order relations in the set of all the probability distributions on a fixed probability space (or, equivalently, of all the random variables on this probability space). These concepts are widely exploited in reliability theory, survival analysis and other problems of stochastic modeling with similar mathematical structure. In some economic applications the term "stochastic dominance" is often in use. Basic definitions and concepts are explained in the monographs by Marshall and Olkin (1979), Shaked and Shanthikumar (1994) and Müller and Stoyan (2002). We deal with two types of orders: so-called usual stochastic order (defined by magnitude of tail probabilities) and a dispersive order (defined by a measure of dispersion of a probability distribution). They seem to be the easiest orders to understand and intuitively interpret also for non-mathematical experts with only basic command in probability and mathematical statistics. Below we give definitions and properties which are useful for analysis of classes of the prior distributions.

Definition 2.1. Let X and Y be random variables on a fixed probability space $(\Omega, \mathcal{M}, P)$ with cumulative distribution functions F and G, respectively. The random variable X is said to be less than Y with respect to the *usual stochastic order* (we write $X \leq_{st} Y$), if

$$\forall u \in \mathbf{R} \quad F(u) \geq G(u), \text{ i. e.} \tag{1}$$

$$P(X > u) \leq P(Y > u). \tag{2}$$

We may observe that Definition 2.1 classifies probability distributions by their tail probabilities, i. e. probabilities of large values. The distribution lower

with respect to this order assigns less probabilities for large values of a corresponding random variable.

It can be proved (see Marshall and Olkin, 1979), that Definition 2.1 is equivalent to the relation

$$E\varphi(X) \le E\varphi(Y) \tag{3}$$

for any nondecreasing function φ such that both sides exist. Observe also that

$$X \le_{st} Y \wedge Y \le_{st} X \Leftrightarrow X =_{st} Y,$$

which denotes equal distributions of X and Y. Further, it is known (Shaked and Shanthikumar, 1994), that the usual stochastic order remains unchanged under summing of random variables and under limit relations.

The usual stochastic order is sometimes called *the first order stochastic dominance*, especially in economic literature.

In Bayesian statistical analysis we deal with a parameter ϑ which is a random variable. Consequently, the sampling distribution of X is the conditional distribution of X under ϑ. Thus the order properties for conditional distributions are needed.

Theorem 2.1. (Shaked, Shanthikumar, 1994). Let X, Y and T be random variables such that the conditional distributions satisfy the following relation:

$$[X \mid T = \vartheta] \le_{st} [Y \mid T = \vartheta]$$

(this means that the definition of the usual stochastic order is satisfied by distribution functions of corresponding conditional distributions) for all $\vartheta \in \operatorname{supp} T$ (the *support* of T, i. e. the least subset $\Delta \subset \Theta$ such that $P(T \in \Delta) = 1$), where π is the probability density function of ϑ. Then $X \le_{st} Y$.

Thus in terms of mixtures of distributions we can say that the usual stochastic order is closed with respect to mixtures and in terms of Bayesian statistic that the usual stochastic order may be transferred to marginal sampling distributions.

Theorem 2.2 (Shaked, Shanthikumar, 1994). Let us consider the family of distributions $\{G_\vartheta : \vartheta \in \Theta\}$, where $\Theta \subset \mathbf{R}$. Let $X(\vartheta)$ be a random variable with the distribution function G_ϑ. For any random variable T with the support included in Θ and with a distribution function F let $X(T)$ denote a random variable with the distribution H defined by $H(y) = \int_\Theta G_\vartheta(y) dF(\vartheta),\ y \in \mathbf{R}$. Let T_1, T_2 be random variables with the common support included in Θ and with

distribution functions F_1, F_2 respectively. Let Y_i, $i = 1,2$, denotes a random variable with the distribution of $X(T_i)$, i. e. with the distribution function

$$H_i(y) = \int_{\Theta} G_{\vartheta}(y) dF_i(\vartheta), \quad y \in \mathbf{R}, \; i = 1,2.$$

If $X(\vartheta) \leq_{st} X(\vartheta')$ for all ϑ, ϑ' such that $\vartheta \leq \vartheta'$ and $T_1 \leq_{st} T_2$, then $Y_1 \leq_{st} Y_2$.

Proof. If $\vartheta \leq \vartheta'$ implies that $X(\vartheta) \leq_{st} X(\vartheta')$, then the probability $P(X(\vartheta) > y)$ is an increasing function of the real variable ϑ for any y. Consequently,

$$P(Y_1 > y) = \int_{\Theta} P(X(\vartheta) > y) dF_1(\vartheta) \leq \int_{\Theta} P(X(\vartheta) > y) dF_2(\vartheta) = P(Y_2 > y),$$

where the inequality is a result of the inequality $T_1 \leq_{st} T_2$ and the condition (3), since $\varphi(z) = P(X(z) > y)$ is a nondecreasing function of the variable z and $E\varphi(T_i) = P(Y_i > y)$. ■

Consequently, Theorem 1 means that the usual stochastic order for the observations (their conditional distributions) under unknown parameters implies the same for their marginal distributions. Theorem 2 means the same under the ordering for the random parameters.

Now we will consider the dispersive order as follows.

Definition 2.2. Let X and Y be random variables with distribution functions F and G, respectively. Let F^{-1} and G^{-1} denote the inverses of the distributions, continuous on the right. It is said that the variable X is less than Y with respect to the *dispersive order* (we write $X \leq_{disp} Y$) iff

$$(\forall 0 < \alpha \leq \beta < 1) \; \mathrm{F}^{-1}(\beta) - \mathrm{F}^{-1}(\alpha) \leq G^{-1}(\beta) - G^{-1}(\alpha). \tag{4}$$

For the above definition we accept

$$F^{-1}(\alpha) \stackrel{def}{=} \inf\{x \in \mathbf{R} : F(x) \geq \alpha\}.$$

An equivalent condition to (4) is that the function $G^{-1}(F(x))–x$ is nondecreasing in x. This is a consequence of the fact that the condition (4) means that the function $G^{-1}(y)– F^{-1}(y)$ is nondecreasing with respect to $y \in (0,1)$. We may observe that the condition (4) classifies probability distributions by a dispersion level expressed by the difference between any fixed pair of quantiles. The below properties may make the sense of the dispersive order more clear.

Properties of the dispersive order (Shaked and Shanthikumar, 1994):

(1) $(\forall c \in \mathbf{R}) \; X \leq_{disp} Y \Leftrightarrow X + c \leq_{disp} Y + c$;

(2) $(\forall a \geq 1) \; X \leq_{disp} aX$;

(3) $X \leq_{disp} Y \Leftrightarrow -X \leq_{disp} -Y$;

(4) If $X \leq_{disp} Y$, then $D^2X \leq D^2Y$ (assuming that both variances exist).

The condition (4) connects the dispersive order with the variance, i. e. the best known measure of dispersion of a probability distribution.

The following theorem gives a relationship between the two order relations.

Theorem 2.3. (Müller and Stoyan, 2002). If the random variables X and Y satisfy the inequality $\inf \text{supp } X \leq \inf \text{supp } Y$, then $X \leq_{disp} Y$ implies $X \leq_{st} Y$.

The proof follows from the fact that then we have for a random variable Z that $\inf \text{supp } Z = \inf\{t : P(Z \leq t) > 0\}$ and from the condition equivalent to (4) given after Definition 2.2.

3. The application to classes of prior distributions

3.1. Conjugate distributions. Let us define a class of distributions by

$$\Gamma = \{\pi_\lambda : \lambda \in \Lambda\},$$

where π_λ denotes a conjugate prior distribution for the sampling distribution with a density $f(x|\vartheta)$. The parameter ϑ may define an order in the set Γ. For some typical distributions the relationships between parameters and orders are as follows (see e. g. Müller and Stoyan, 2002):

- the gamma distribution *Gamma*(λ,p) with the density proportional to $t^{p-1}e^{-\lambda t}$: increasing p follows increase with respect to the usual stochastic order and increasing λ follows the decrease in this order; moreover, for p=1 this decrease happens also for the dispersive order;
- the beta distribution *Beta*(p,q): increasing p follows increase with respect to the usual stochastic order and decreasing q follows decrease with respect to it;
- the normal distribution $N(\mu,\sigma)$: increasing μ follows increase with respect to the usual stochastic order and increasing σ follows increase with respect to the dispersive order.

It implies useful consequences, i. e. order relations may be transferred onto posterior distributions. If a sample distribution comes from a natural exponential family, i. e.

$$f(x|\vartheta) = C(\vartheta)h(x)e^{\vartheta x} = h(x)e^{\vartheta x - \psi(x)},$$

where ϑ, $x \in \mathbf{R}^k$, then the conjugate prior distribution of the parameter ϑ has the form

$$\pi(\vartheta \mid \mu, \xi) = K(\mu, \xi)\mathrm{e}^{\vartheta\mu - \xi\psi(\mathrm{x})}$$

and it is defined on the so-called natural parameter space

$$N = \left\{ \vartheta \in \mathbf{R}^k : \int_{\boldsymbol{X}} e^{\vartheta x} h(x)dx < +\infty \right\},$$

with $K(\mu,\xi)$ being a norming constant and $\mu \in \mathbf{R}^k$, $\xi \in \mathbf{R}$ are parameters of the prior distributions (what means that the "joint" parameter space has the dimension k+1). Then the posterior has the form $\pi(\vartheta \mid \mu + x, \xi + 1)$ (see e. g. Robert, 2001). It follows that if the parameter "rules" over a stochastic order, then the order is transferred onto the posteriors. This implies that bounds for parameters may have deeper meaning.

3.2. Distributions with ε-contamination. Let us define a class of distributions by

$$\Gamma_\varepsilon = \{\pi : \pi = (1-\varepsilon)\pi_0 + \varepsilon q, q \in Q\},$$

where Q is a class of prior distributions defined by density functions, and the number $\varepsilon \in (0,1)$ is fixed. There exist results on ordering of such mixtures under variable ε, but they have no meaning for classes of prior distributions, maybe apart from infinitesimal reasoning as $\varepsilon \to 0$. For any fixed ε the usual stochastic order in Q evidently defines such an order in the class Γ_ε and so does in the set of corresponding marginal sampling distributions.

In the considered case the posterior has the following form:

$$\pi(\vartheta \mid x) = \frac{(1-\varepsilon)m_{\pi_0}(x)\pi_0(\vartheta \mid x) + \varepsilon m_q(x)q(\vartheta \mid x)}{(1-\varepsilon)m_{\pi_0}(x) + \varepsilon m_q(x)} =$$

$$= \lambda(x)\pi_0(\vartheta \mid x) + (1-\lambda(x))q(\vartheta \mid x),$$

where

$$\lambda(x) = \frac{(1-\varepsilon)m_{\pi_0}(x)}{(1-\varepsilon)m_{\pi_0}(x) + \varepsilon m_q(x)}$$

and $m_\pi(x) = \int_\Theta f(x \mid \vartheta)\pi(\vartheta)d\vartheta$ is the marginal sampling distribution under the prior π. We can expect that the transfer of an ordering of priors onto posteriors is very doubtful here. The following counterexample shows that the doubt is justified.

Example. Let us consider the normal distribution $N(\mu, \sigma^2)$. The parameter σ determines the dispersive order. Let $X \sim N(\mu, \sigma^2)$, let π_0 be the density of

$N(\mu,\sigma_0^2)$ and q be a density of $N(0,\tau^2)$, where $\tau > 1$. The posterior is not normal but its variance is known to be

$$D_\pi^2(\vartheta \mid x) = \lambda(x)D_{\pi_0}^2(\vartheta \mid x) + (1-\lambda(x))D_q^2(\vartheta \mid x) + \lambda(x)(1-\lambda(x))\times$$
$$\times\left[E_{\pi_0}^2(\vartheta \mid x) - E_q^2(\vartheta \mid x)\right]^2.$$

For fixed σ, μ, σ_0 and x this is a function G of the variable τ. If G is not monotone for some fixed values of the above parameters, then this contradicts the dispersive order for the posterior distributions. Let σ=1, μ=0, $\sigma_0 = 2$. We obtain that for some x the function G is not monotone. For example if $x \geq 2.2$ then for $\tau \in (1.0, 8.0)$ the function G has at least one local minimum point more than 1.1.

3.3. The class of priors defined by generalized moments. Let

$$\Gamma = \left\{ \pi : \int_\Theta H_i(\vartheta)\pi(\vartheta)d\vartheta \leq \alpha_i,\ i = 1,2,...,k \right\},$$

where H_i are real functions on Θ such that the integrals exist. For example if we substitute H_i by ϑ^{r_i}, we obtain the r_i th moment, for $f(x_i \mid \vartheta)$ – a value of the marginal sample density, for $\mathbf{1}_{C_i}(\vartheta)$, $C_i \in \mathcal{F}$ –prior probability of a set C_i, for $\int_{A_i} f(x \mid \vartheta)dx$, $A_i \in \mathcal{M}$ – marginal sample distribution probability of a set A_i.

For nonnegative random variables the usual stochastic order implies the *moment order*, defined as follows:

$$X \ \pi \ Y \Leftrightarrow (\forall i \in J)\, EX^i \leq EY^i,$$

where J is the set of such positive integers that the ith moments of X and Y exist. Besides, marginal sample distributions preserve the usual stochastic order of prior distributions (Theorem 2.2). In such a way, taking the above four examples of generalized moments, the first one means that the usual stochastic order of prior distributions of nonnegative parameters implies inequalities for moments and the third and the fourth ones mean the analogous implication for the sets C_i, A_i (respectively) having the tail form.

3.4. Mixtures of distributions. Let us define the following class of distributions for W being a fixed set of weight functions:

$$\Gamma = \left\{ \pi : \pi(\vartheta) = \int_S \pi_s(\vartheta)w(s)ds,\ w \in W \right\}.$$

Then the usual stochastic order in W implies the same one in Γ which is transferred onto the set of marginal sample distributions.

3.5. Distribution band. In this section we identify distributions with its cumulative distribution functions. For cumulative distribution functions F_L and F_U of fixed prior distributions we define a class of distributions (cumulative distribution functions) as follows:

$$\Gamma = \{\Pi : (\forall t \in \mathbf{R})\, F_L(t) \le \Pi(t) \le F_U(t)\}.$$

Such a set is called the *distribution band*. Basu and DasGupta (1995) obtained results for extreme values for posterior functionals if the prior distributions run over such a class. Moreover, for the Kolmogorov metric in the space of probability distributions, which is defined as

$$\rho_K(F,G) = \sup_{x\in\mathbf{R}} |F(x) - G(x)|,$$

it is possible to see that the neighborhood of a fixed distribution F_0 with a radius ε has the form of the distribution band with $F_L(t) = \max\{F_0(t) - \varepsilon, 0\}$ and $F_U(t) = \min\{F_0(t) + \varepsilon, 1\}$. The distribution functions in the bounds may be defective, i. e. the limit in $+\infty$ less than 1 for F_L and the limit in $-\infty$ more than 0 for F_U. For such neighborhoods the extremes of posterior functionals was given by Boratyńska and Zieliński (1993) and Boratyńska (1994).

4. Concluding remarks

Applications of the simplest stochastic orders to defining classes of prior distributions in Bayesian robustness problems have shown to be difficult. The dispersive order which seemed to be intuitively interesting has been inadequate for the problem apart from a very simple (important, however) case of conjugate distributions. More relationships to typical definitions of classes of priors have been shown for the usual stochastic order and the most natural class, i. e. the distribution band was thoroughly investigated about ten years ago.

REFERENCES

S. BASU, A. DASGUPTA (1995): Robust Bayesian analysis with distribution bands, *Statistics & Decisions* 13, 333—349.

J. BERNARDO, A. F. M. SMITH (1994): *Bayesian Theory*, J. Wiley, Chichester.

A. BORATYŃSKA (1994): Infinitesimal robustness in Bayesian statistical models (in Polish), *Matematyka Stosowana* XXXVII, 67—106.

A. BORATYŃSKA, R. ZIELIŃSKI (1993): Infinitesimal Bayesian robustness via the Kolmogorov metric, *Applicationes Mathematicae* 22, 139—143.

A. W. MARSHALL, I. OLKIN (1979): *Inequalities. Theory of Majorization.and Its Applications*, Academic Press, New York.

M. MĘCZARSKI (1998): *Robustness Problems in Bayesian Statistical Analysis* (in Polish), Oficyna Wydawnicza SGH, Warszawa.

A. MÜLLER, D. STOYAN (2002): *Comparison Methods for Stochastic Models and Risks*, J. Wiley, Chichester.

D. RÍOS INSUA, F. RUGGERI (Editors, 2000): *Robust Bayesian Analysis*, Springer, New York.

C. P. ROBERT (2001): *Bayesian Choice*, 2nd edition, Springer, New York.

M. SHAKED, J. G. SHANTHIKUMAR (1994): *Stochastic Orders and Their Applications*, Academic Press, New York.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1131—1150

TOWARDS FURTHER DEVELOPMENT OF THE POLISH CONSUMER PRICE STATISTICS

Alina Gluchowska[1]

ABSTRACT

The article summarises surveys on retail and consumer prices which used to be carried out by the official statistics in Poland and it describes also main aspect of the methodology of consumer price indices applied currently in the country, i.e. at the time of the 85th anniversary of the Polish Central Statistical Office (GUS). Respecting the long history of statistical surveys on prices paid by Polish households, it has to be said that the biggest step was made in the second part of the nineties when the EU harmonised methodology was gradually implemented in GUS. Then Poland joined several other European countries covered by harmonisation of consumer price measurement.

International methodology is developing constantly and therefore further improvement of consumer price statistics is needed. Development in this area of statistics in Poland should take advantage of recent international achievements of price statistics, such as the new EU legislation on HICP, the CPI Manual (ILO, 2003) and the 2003 ILO resolution on CPI. The article illustrates progress made in the Polish CPI and indicates those aspects of it which need further improvement.

In view of EU membership it seems reasonable to adopt the Eurostat HICP methodology also for the Polish national CPI. The economic aspect of the survey is of importance here. Having in mind careful management of resources it seems to be justified to apply in future some of the methodological principles, namely those relating to sample selection and price recording, as common to both surveys. The differences concerning some other aspects: system of weights, coverage, product scope, aggregation and data publication, could be formulated so as to meet the needs of the main users of data.

Key words: consumer price survey, HICP, inflation, price collection, weights.

[1] Kent Consultancy Ltd., Canterbury, England. On leave from Central Statistical Office, Social Statistics Division, Warsaw, Poland. E-mail: a.gluchowska@op.pl.

1. Introduction

Changes in prices of goods and services purchased by the population for consumption have been subject to statistical surveys for a few centuries. The results of these surveys are regarded as key statistical data, are required by all kinds of users on both the macro- and microeconomic level as well as by individual consumers.

Among the available Polish sources of data on prices it is necessary to mention the Concise Statistical Yearbook of Poland, published in London in 1941, which reported the cost of living indices for workers' families in Warsaw since 1924 and retail price indices for Warsaw since 1929.

Price collection was undertaken again in the first years of the post-war history of Polish statistics and retail price indices were published for 1950. The description of survey methodology and practice in retail price collection in the following decades (up to 1990) can be found in the publications by Kordos and Dublanka (1982), Kordos and Kowalski (1983), Luszniewicz (1984), Kordos and Czajka (1987), Kordos (1988), Nowak (1992) and Toczynski (1989) as well as in the CSO studies on methodological rules (1985). These works discuss the measures taken to ensure reliable information on retail price dynamics in the situation of shortages of goods, black market economy, and hidden price rises followed by repeated significant price increases at the beginning of the transition period between the eighties and nineties.

The demand for high quality statistical data reflecting the economic and social conditions is a stimulating factor for methodological development in the field of price index calculation. It can be pointed out that the development of consumer price measurement methodology became much faster starting from the late eighties. For instance, in 1987 the International Labour Organisation (ILO) issued a resolution on consumer price indices based on the latest developments in the theory of statistics. It was the fourth resolution by ILO in this area and it happened at the right historical time for the improvement of national methodologies of price measurement, especially due to the fact that the previous resolution had been adopted by ILO 25 years before. The recommendations given in 1987, reported for Polish readers by Kordos (1988), provided thorough grounds for methodological changes to be introduced by the Polish Central Statistical Office (CSO). The implementation programme was published in 1989 by Toczynski, who was at that time the director of the newly established Price Statistics Division of GUS. During the following years the methodological principles of CPI measurement were improved, among others following the detailed instructions for CPI measurement given in the manual published at the initiative of ILO and edited by Turvey (1989).

The Polish CSO started to modernise retail price measurement in the early nineties, basing its concepts on both the international experience and domestic forum discussions. Statistical methods of price surveys were the main topic of the

scientific conference organised in 1989 by the Polish Statistical Association, the Committee on Econometrics and Statistics at the Polish Academy of Sciences and CSO ("Statistical News Library" 1990). In 1991 the CSO organised an international seminar on price statistics with the participation of experts from USA (Bureau of Labor Statistics — BLS) and Eurostat.

The adjustment of Polish price statistics to international standards was carried out in close co-operation with international organisations, especially Eurostat. The description of methodological changes introduced in the first half of the nineties was presented in several articles such as those by Czajka and Nowak (1991), Nowak (1992), Turek (1992,1995), Turek and Gluchowska (1995).

In the early nineties the CSO established co-operation with Eurostat where since 1993 intensive work has been carried out under the guidance of J. Astin (1999) with the aim of creating a conceptual and legal framework for harmonised indices of consumer prices (HICP). A number of legal acts were passed imposing on EU member states an obligation to apply certain methodological rules. A few years later, Eurostat started to publish monthly reports on consumer price indices for the EU member states, calculated in accordance with the comparable harmonised principles. In the second half of the nineties the programme of harmonisation also covered the EU candidate countries so that since 1999 they have submitted preliminary data on HICP to Eurostat. Then, together with the countries outside the EU which produce HICP (Iceland, Norway and Switzerland), European harmonisation of consumer price measurement covered some 30 countries. The Polish price statisticians follow current trends in methodology suggested by expert groups, such as the international group on price indices appointed in 1994 by the UN Statistical Commission, known as the "Ottawa Group" which gathers every 15-18 months to initiate discussion on difficult issues and promote the best solutions. Mention must also be made of the report by the Boskin Commission (1996), which revealed many weak points in the US CPI measurement methodology. It provoked international discussion which stimulated development of the methodology, especially in the area of monitoring potential sources of error including those brought about by disregarding changes in product quality. As a result of the research inspired by the Boskin report and conducted in the following years by a specially appointed team of experts from American universities, a comprehensive methodological publication (2001) presented the academic standpoints concerning difficult aspects of consumer price and cost of living measurement.

In summary, the last decade of the 20th century was marked by further integration of international actions towards improvement of the CPI methodology. However, methodological principles still vary from country to country, which affects comparability of results. Work is continuing to find solutions for the controversial methodological problems discussed by price statisticians.

2. Definition of CPI, its main purposes and related terminology

The definition of an index depends on the purpose for which it is intended. In general, the consumer price index (CPI) is a measure of the changes over time in the prices of goods and services purchased by the population for consumption. However CPIs are used in practice for different purposes, in the great majority of countries they are treated as the main inflation measure. The harmonised index of consumer prices (HICP) plays a similar role.

Although some countries (USA, Sweden) base their CPIs on a similar concept to that of the cost of living, in practice it is not possible to design directly a true cost of living index as an indicator which measures the change between two periods in the minimum expenditure which would be incurred by a utility-maximising consumer, whose preferences or tastes remain unchanged, in order to maintain a given level of utility (standard of living) in both periods. It would be equally difficult to measure an ideal whole-economy inflation rate covering transactions in all the sectors of economy (though some countries are trying to make estimates in this area). Therefore, the CPI is considered as the closest to the measure of both price and cost of living changes. In Poland, the CPI is an indicator reflecting inflation in consumer prices as its methodology is based on the Eurostat HICP approach.

There is a strong connection between the definition and objective on the one side and the name of the index on the other. The term designating the index has changed in Poland over time. In the first half of the twentieth century the term "cost of living indicator" was the most popular. However, after the Conference of European Statisticians in 1947 "consumer price index" was recognised as more adequate and the former name was gradually dropped. Nevertheless, in Poland up to the late eighties price indices for particular groups of households were referred to as “cost of living indices”. Since 1991 the term "index of consumer goods and services" has been used. The term "consumer" was for the first time applied to goods and services purchased by the population so as to indicate that they were designated for consumption, and not for production or investment. Before 1991 two types of indicators were measured and published by the Polish CSO - the previously mentioned cost of living index and the retail price index. The first was calculated for different groups of households using household expenditure on consumer goods and services as the weighting system; the other for the total population using the retail sales structure as weights. After major changes in the price measurement methodology in 1991, data on household expenditure based on household budget surveys (HBS) have been used as the main but not the only source for designing the weighting systems for consumer price indices.

The CPI is commonly referred to as the inflation rate, especially when measured over a 12-month period. It should be remembered, however, that this is a simplification, because this indicator is just one of several possible inflation

measures. Inflation is understood as a simultaneous, persistent increase in the general level of prices. It might affect all the economic spheres of the country.

The basic terminology of price measurement is contained among others in the EU legislation on HICP (2001), and material prepared after the meeting of the expert group appointed by ILO in 2001 and in the ILO Manual (2003). Included in the terminology are:

- Basket — a term used often in two meanings: as a list of goods and services chosen for the survey, whose prices are observed by price collectors; or as a weighting system (quantities) adopted for price index calculation in the surveyed year;
- elementary aggregate — the most detailed grouping of goods or services (weights), for which household expenditure is available for a sufficiently long period (e.g. a year);
- price survey area — geographical location within the country, where price observations are made;
- outlet — a shop, a supermarket, a department store, a pharmacy, a petrol station, a service point, a permanent market-place and other place of sales where prices may be observed (e.g. internet),
- representative item (or product) — a good or service offered in an outlet, whose price is under observation;
- consumption — denotes all goods and services, which are bought by consumers for purposes other than production or investment. Consumer expenditure may refer to any of the following transaction stages: „acquisition", „consumption" or „payment";
- coverage of the index — relates to socio-economic groups of population, regions, products and outlets;
- index reference period — time point or a period (e.g. month), on which the index is based to; often denoted as „100" or „1".

The Polish CSO uses this terminology and makes it more familiar to the data users publishing it in its statistical internet glossary.

The adoption of common terminology in price statistics, and its implementation in statistical practice at the international level, is one of the conditions of achieving international comparability.

3. Sampling

Price surveys on consumer goods and services are conducted on the basis of a sample which covers price survey areas, outlets and representative items. The virtues of random sampling are well known; though it has to be said there are practical difficulties for its implementation in the CPI. But it is not possible in practice to include in a survey all existing products and outlets. For these reasons both Eurostat and ILO tolerate surveys based on purposively chosen samples.

This might lead however to bias resulting from subjective choices, and attempts should be made to gradually implement random sampling methods for price surveys. In the Polish CSO, focusing more on the other crucial CPI methodology aspects, no such trials have been undertaken. A prerequisite for this work is the existence of good quality (updated) registers at each stage of sampling, namely, price collection areas and outlets. Dynamic development of the market and the retailing structure in Poland is also a serious obstacle.

Rapid changes in the Polish economy, in particular the structure and volume of supplies together with sharp upward price movements, which began in 1989, required a radical increase in the sample for retail price surveys. The list of goods and services for which retail prices were collected for the CPI needs was at that time expanded from 1,200 to about 1,500. The number of areas surveyed was increased to 307 (as compared with about 120 previously). In the following years the sample size was determined by the CPI accuracy requirements conditioned by such factors as: measurements for all socio-economic groups of households (since 1994), classification changes (in 1999) and data breakdown according to the new administrative division of the country (16 voivodships since 2000).

In 2004 prices of about 1800 consumer goods and services were collected in 308 survey areas. The criteria used to determine the number of price collection areas in a given voivodship included the locations and density of retail outlets relative to the size of the local population, as well as sales values and volumes.

The number of representative goods and services, price survey collection areas and frequency of collection have an effect on the number of prices observed during a month, which for Poland amounts to about 0.5 million. To compare, in the EU member states the number of representative items[1] ranged in 1997 between 263 and over 16,000 and that of prices observed during a month - from 6,700 to over 320,000. The sample size is not always dependent on the territory of the country and the differences may result from the level of detail in describing the representative items. The Polish CPI sample seems to be quite large in comparison.

As the inflation rate remains at a low level during recent years and may be expected to be stable also in the future, this is the right moment to review the sample for the Polish CPI. When considering rationalisation of the sample size aimed at a reduction (starting e.g. from the number of price collection areas) it should be remembered that the sample needs to be large enough to provide, for example reliable regional indices and accurate product average prices for those users who require them.

[1] The data are related to the sample size adopted by EU member states for HICP since 1997. At present, the sample sizes may be slightly different because of an increase in the HICP coverage since 1999. The number of observations is usually given as an estimate (as it is in Poland), assuming that all the representative items are recorded in every survey area.

4. Frequency and timing of price collection

Following international recommendations prices should be collected once a month but there are no regulations limiting more frequent observations, especially when prices of certain products might change frequently and significantly during a month or when it is in accordance with users' requirements and there are no financial obstacles. According to the law, Eurostat allows for quarterly price collection on products the prices of which are known to change infrequently (e.g. rents in some countries).

The ILO and EU guidelines do not specify the day of the month for price collection. On the other hand, HICP principles provide a requirement to prevent systematic skipping of price reductions and promotions on many products, because this could affect the survey results.

In Poland, starting from 2004, prices are collected once a month, except for vegetables and fruit for which prices are collected three times a month. The schedule for price collection is made by regional statistical offices. In the "Rules of retail price collection " issued by the CSO for price collectors (2002 edition updated) it is pointed out that the dates of observation for a given group of goods in a given area should be carefully chosen to avoid e.g. a concentration of collection at the beginning of the month and then at the end of the following month. In accordance with the ILO recommendations, a fixed interval should be kept between month to month price collections. The spread of observations over time in different price collection areas provides information about average monthly prices. The ILO Manual says that price collection days (and sometimes times) should be set in advance. This is followed in Poland, but given that the price collection timetable is currently rather fortuitous, it would be advisable to establish a more fixed frame of price collection for the country — especially when the sample reduction is considered. Using international recommendations in this area, schedules might be drawn up for different price survey areas or for particular type of products, for example , food — the first week of the month, housing — the second week, clothing — the third week, other goods and services — the fourth week. Such a scheme should take careful account of, for example frequency of changes in prices which usually are introduced for particular group of products. Although this is as simple as arbitrary, it would introduce a degree of consistency and order into the survey. EU legislation, expected soon, on this issue will also be helpful.

5. Price collection techniques

In most countries price measurement is carried out by price collectors, but only in some of them (like Poland) are they full-time employees of statistical offices. Statistical offices sometimes commission this task to specialised agencies.

Various methods and techniques of data collection are applied, depending on the type of prices and data sources.

For prices fixed administratively, price lists and administrative decisions issued on the national scale are used as the source. In such situations the information on prices is collected centrally. Another type of central collection is the use of catalogues published by large retailers. Telephone and fax are also used centrally for certain items, particularly services. These techniques are applied by the Polish CSO.

Among other price collection techniques suggested in the ILO Manual there are also internet sites, which are more and more commonly used as a source of information on price changes. In Poland the internet is mostly treated as one of the sources of information for data validation. Some countries are experimenting with scanner data (the data recorded electronically when consumers pay for their purchases at check-outs). Statistical offices also conclude agreements with large retailers which report detailed statistics on their prices and turnover.

The most common technique however remains is the questionnaire in paper form, though more and more countries are providing their price collectors with portable computers, whose main advantages are improving comfort of work, shortening the time of price collection and helping to avoid errors due to simultaneous automatic data checking. In such countries as the UK, Finland, France and Russia, these types of computer are being used successfully by price collectors. The Polish CSO is currently also introducing portable computers for price collection.

6. Price and quality change

The treatment of quality change as one of the factors influencing the price survey results has been discussed for many years, but with a greater intensity since 1996 after the publication of the Boskin report. Experts are not united in their opinions on the range of corrections to be made with regard to quality change, but most admit that it is a very important methodological aspect. There are more and more publications available, describing new methods of quality adjustment. The two types often indicated are implicit and explicit adjustments. This issue has already been referred to in several Polish publications on this topic e.g. by Smilowska (2001), Gluchowska (2002) and Smilowska and Cynkier (2002). The Polish CSO has been using mostly implicit methods so far. However, experiments should be continued to develop quality adjustments appropriate for particular groups of goods and services. Among the quality adjustment methods most frequently suggested by international bodies are:

- overall or class mean imputation,
- overlap or bridged overlap,
- link-to-show-no-change,
- essentially equivalent substitution,

- judgmental methods,
- quantity adjustment,
- differences in production costs or option costs,
- hedonics.

However, considering that some of the above methods have been much criticised during international meetings and conferences (the method chosen might affect the survey results) it would be important for international harmonisation to have the permissible methods determined and to design guidelines for their practical use. Adaptable international standards in this area have not been developed so far.

7. Formula and weighting system

The Laspeyres formula is the one most commonly applied for CPI measurement, though it is not often used in its classical form, mainly because of rapid and significant changes in retail markets which do not allow for a distant base period. Instead, the so-called modified Laspeyres index formula is usually applied.

At the level below the elementary aggregate, geometric means of price relatives are applied in Poland (Jevons price index type). At higher levels of aggregation, for practical reasons the modified Laspeyres index formula is adopted in Poland. According to this formula the system of weights is taken from the base period and not the survey period. The base period for the Polish CPI, both for weights and prices, is the previous year. In HICP measurement the EU regulations allow the use of weights based on the annual structure of household consumer expenditure relating to a period not earlier than 7 years before the current year, but updated every year with new groups of goods and services if they are purchased by consumers in sufficient quantity (for HICP at least 0.1% of total consumption). Taking into account continuing changes in household consumption patterns and development of the market, Poland updates the weights annually both for CPI and HICP measurement.

When designing the weighting system it is necessary to pay particular attention to certain kinds of products, such as seasonal goods and services, insurances, second-hand goods, consumers' expenditure abroad, expenditure of institutional households, interest, own consumption as well as housing-related expenditure (on purchase, reconstruction and decoration of houses and flats). According to the ILO resolution it is an optional decision of national statistical offices whether to include these items in the CPI or not. The current EU regulations on HICP are stricter. Following them, out of the groups of products mentioned above the following should be included in the index: seasonal goods and services, non-life insurance (it is recommended to take the mean value from the last three years for weights), second-hand goods (in particular cars sold by companies), foreigners' expenditure within the country and expenditure of

institutional households[1]. The items to be excluded are: households' expenditure abroad, interest on credit and own consumption. The Polish CSO introduced these rules for HICP. In case of CPI the latter two are not included.

Methodological work carried out recently by Eurostat has been aimed at including in HICP the expenditure on purchasing and building dwellings. These items would be treated similarly to other durable goods, such as cars. The Polish CSO participated in the pilot survey on this topic. As proposed by Eurostat, the price index for owner occupied housing would be in the first stage calculated as an index accompanying the HICP. In some countries such expenditure is included in the national CPI, though methods applied for index calculation vary. In Poland, until 1998, the CPI included the so-called “cost of acquisition of a flat”, which was excluded from the national CPI when HICPs began to be calculated. According to the definition adopted by the national accounts and household budget surveys, the purchase of a flat or a house is not treated as consumer expenditure. The new methodology in Poland for the owner occupied housing price survey has to be developed and introduced following any forthcoming the EU regulation.

International recommendations stipulate that another source of data for the weighting system for CPI, in addition to household budget surveys (HBS), is provided by the national accounts statistics. So far HBS is treated in the Polish CSO as the basic source of data for the weighting system construction. Additionally, some information from the national accounts and trade statistics is used. HICP methodology is based on the concept of "Household final monetary consumption expenditure". Current international trends in statistics and the EU requirements in the scope of HICP also point to the possibility of using national accounts statistics as the main source of data at the higher levels of aggregation and HBS results at the lower levels.

As for seasonal goods, statistical practice allows the use of both variable weights (depending on monthly consumption), i.e. the so-called Rothwell method (used in Poland since 1992) and fixed annual weights. There are different opinions on the application of these methods. Both kinds of weights — fixed and variable — have advantages and disadvantages, and as a result of the lack of uniform international recommendations both methods are accepted, even though they are likely to produce differing results.

Further studies are also required to consider the additional inter-voivodship weighting level, for both the national CPI and HICP purposes. So far for measuring both types of indices the weighting has been based on the number of price observations which, by definition, is proportional to the size of the voivodship.

---

[1] Like welfare houses, houses for the elderly, prisons etc.

8. Range of consumer price indices

The various types of CPI calculated should be determined by the requirements of statistical data users and the objective of CPI measurement. In general, European (EU) and other international (UN ECE, ILO) recommendations concerning the geographic, demographic and product coverage of CPIs are convergent, but some differences can be found in the detail.

As for the geographic coverage, CPIs fulfilling the conceptual criteria of an inflation measure should cover the goods and services accessible on the territory of the country, satisfying the consumer needs of the population of that country (domestic concept). This means in particular that the sample includes both domestic products and imported ones sold in domestic trading outlets, with the exclusion of citizens' expenditure abroad. The CPI measured by the Polish CSO fulfils these criteria.

In accordance with the EU regulations HICP measurement requires taking into account foreigners' expenditure (excluding business expenditure) made on the economic territory of the country. Therefore, the Polish CSO has made estimates needed for HICP, based on the data available, i.e. from the pre-border survey carried out until recently and from the national accounts. However, the data are not available in the proper aggregation or in sufficient detail. This question should be carefully examined in the near future, particularly regarding the possibility of obtaining data from other potential sources such as tourist companies. It is supposed, however, that in Poland the share of foreigners' expenditure in the total expenditure of households is not as significant as in many EU countries, where the tourist inflow is much greater. The weighting system used for the Polish national CPI does not take account of the estimated expenditure of foreigners, mainly because this index is used for the indexation of pensions and retirement payments in the country as well as the deflation of other domestic economic data. But more attention should be paid in the Polish CSO to the inclusion of foreign tourists and residents whose numbers seem likely to increase in the future — in the same way that other EU member states include for example foreigners working in the international institutions.

In accordance with the international recommendations concerning geographic coverage, the CPI should be representative of the whole country, though its breakdown by region is not required. The decision on this question is taken individually by national statistical offices with respect to country size, the needs of data users and funding available. As a result, only some countries (e.g. Germany) produce regional CPIs. Most EU countries publish a CPI for the whole territory of the country and possibly for certain areas or cities. The Polish regulations oblige the CSO to present a CPI by voivodship every quarter, while the HICP is determined as a national index only, without territorial breakdown.

Following international recommendations concerning demographic coverage, the CPI should cover all types of household, regardless of their

economic status. However, the updated resolution of ILO (2003) is more liberal than EU regulations (2001), in its statement that for practical reasons (such as difficulties in reliable data collection) or because of financial problems the expenditure of certain types of households, e.g. very rich or very poor ones or those living in certain regions of the country could be excluded from the weighting system. It is only necessary to inform the data users about it. The Polish CPI takes into account all types of households regardless of their economic status.

Since January 2003 the Polish HICP in accordance with the regulations has included the estimates of institutional households' expenditure based on national accounts data. However, considering the fact that these figures are not at the necessary level of detail, this issue requires further study and improvement of the estimation method. CPI for domestic needs does not include institutional households.

EU regulations do not require HICPs for particular types of household. However, the ILO resolution indicates that CPIs are produced for particular population groups if such data are needed in a country. Since 1993 the Polish CSO has published CPIs for seven groups of households[1]: employees and employees-farmers (combined), employees, employees-farmers, farmers, the self-employed, retirees and pensioners, and households maintained from non-earned sources. In addition, from 1998 separate indices are calculated for retirees and for pensioners. These indices reflect price variations on consumer goods and services purchased by particular socio-economic groups. The CPI for the combined group of employees and employees-farmers is used for the calculation of real wages and that for retirees and pensioners — as auxiliary information for determining the valuation scale of retirement payments and pensions.

In accordance with the recommendations concerning product coverage, CPIs should include goods and services purchased by households in monetary transactions in order to satisfy their consumers' needs (excluding investments and production). This indicates that the so-called own consumption (e.g. products from own garden or farm) as well as goods and services received as gifts or donations or part of consumption provided by the state is not taken into account by the Polish CSO .

Most countries, including Poland, take into account only those prices which are paid in legal transactions. This is the reason why such products as narcotics are not included in the Polish CPI. However, it is recommended to take into account hidden economy transactions, if possible. The indices calculated by the Polish CSO include e.g. the prices of lessons given by private tutors. It would be worthwhile to review the range of hidden economy transactions (using the

[1] Before 1993, CPIs were produced by CSO for the four socio-economic groups of households (employees, employees-farmers, farmers, pensioners and retirees) as well as by other aggregations, e.g. employees' households by selected occupational groups.

national accounts definition) to identify the groups of products the prices of which should and could be observed.

The prices collected for CPI measurement should include all kinds of indirect taxes, and perhaps also some direct ones if they represent actual consumer expenditure. This principle equally applies to the prices of goods and services subsidised by the state, and in such cases the index should again account for that part of price which is actually paid by the consumer. The Polish CSO follows these rules, for example in the case of medicine prices (partially subsidised by the state). Similarly, another standard implemented by the Polish CSO for ensuring the CPI includes only actual prices paid by the consumer, is the use of prices reduced due to promotions, post-seasonal price reductions etc.

The survey should include the prices paid in trading outlets (usually non-negotiable), also those agreed between the seller and the buyer (e.g. on the market place) as well as those set administratively. The Polish survey covers all types of prices mentioned above; including those set administratively, referred to as official and controlled prices. The scope and definition of these prices vary from country to country. Recent investigations made by the European Central Bank (ECB) point out the need to carry out more thorough studies. National variations in financial and budgetary systems, also different tax systems, may bring as a result a lack of harmonisation in this area.

The index should cover, in the appropriate proportions, the goods and services purchased in different kinds of trading outlets (mentioned in the first part of the article). The data for this purpose should basically be derived from trade statistics. The trade statistics properly developed seems a significant factor for the future of many statistical surveys (not only for price measurement purposes). Trading outlets include both small "corner" shops usually selling everyday products, and also specialised ones with a narrow range of products as well as large supermarkets, department stores, pharmacies, petrol stations etc. Price collectors should record prices not only in average standard outlets, but also in luxury shops, trademark shops and fixed markets. These rules are implemented by the Polish CSO, though the choice of trading outlets is left to regional statistical offices which not only have access to local trading reporting but also have gained empirical knowledge about the local market.

9. Classification

In 1999 the Polish CSO adopted uniform classification standards, starting to use the Classification of Individual Consumption according to Purpose (COICOP) as recommended by international organisations. It was applied earlier (1997) for HBS. The classification was worked out and published by UN (the latest edition of 2000) and it is used by many countries throughout the world. Eurostat adapted COICOP to HICP purposes and this version is known as COICOP/HICP. In Poland, as in many other countries, the same classification is used for the national

CPI. COICOP/HICP consists of 12 divisions (two-digit levels), 39 classes (three-digit levels) and 93 groups (four-digit levels). Further breakdown below the COICOP groups is recommended by Eurostat, but it depends on the availability of statistical data at an appropriate level of detail. Further levels of division (five-, six- and seven-digit ones) are applied by the Polish CSO. A more detailed breakdown depends on the nature of a given group of products i.e. assortment and price differentiation.

Before 1995 the following classification standards were applied by CSO for retail price surveys: NAS (Nomenclature of Articles and Services for Consumption Survey) and ICGS (International Classification of Goods and Services for Households). Considering the fact that there are significant differences in aggregation between those two nomenclatures and COICOP, no transition key was worked out or retrospective data estimated. Therefore, continuity of data on consumer prices is ensured only for overall CPI and for certain aggregates (back to 1950 for annual data and to 1982 for monthly data).

10. Data validation

The literature on price statistics gives a classification of possible errors which may affect CPIs. In addition, the ILO resolution reminds price statisticians of certain aspects of the survey which can become the source of errors. They include such factors as incorrect selection of representative goods and services to be surveyed, mistakes in price recording, disregarding quality changes, lack of sample updating with new goods and services, inappropriate calculation formula. Some of the negligence that can affect CPI measurement was described by Kordos (1990). These errors can be avoided by applying strict methodological principles, observing them carefully and checking the data. The Polish system of price statistics has a tradition of data validation (also automatic with the use of special software) at several stages of the survey: preparation and price recording in regional statistical offices and index calculation in the CSO.

Following international standards, expert committees are appointed in many countries, involving researchers, statisticians and main users of price data, such as trade unions, central bank, ministry of finance etc. The task of such a group is to formulate, analyse and accept the methodological principles. In Poland a similar role is played by the Statistical Council, the Scientific Council of Economic and Statistical Research Establishment, and the Programme and Methodological Commission. However, no special committee has been appointed for consumer price statistics.

11. CPI types, their purposes and data users

There are three basic functions to be performed by consumer price indices: as a deflator, in economic and market analysis and for indexation. In order to comply with them according to the public statistics programme the following types of indices are produced by the Polish CSO:

a) price indices of consumer goods and services (CPI):
   - as overall CPI and by the aggregates required by users — computed annually since 1950 and monthly and quarterly since 1982. The scope of indices is defined by the analytical and deflation requirements of the data users: both internal (mainly national accountants) and external (Ministry of Finance, Central Bank, IMF, ILO, World Bank, OECD etc.) The indices are also used for indexation purposes in accordance with national legislation;
   - for particular socio-economic groups (as mentioned earlier in this paper). These indices are produced monthly, quarterly and annually. They are published as an overall CPI for each of socio-economic groups and for aggregated product groups. These indices are used for indexation and deflation purposes;
   - by region — since 2000 a CPI is published quarterly and annually for each voivodship and for aggregated product groups. Such data are used for indexation purposes in accordance with the legislation mostly on building cost estimation;

b) harmonised indices of consumer prices (HICP) — calculated since 1997 and transmitted by CSO to Eurostat in accordance with unified methodology, to be used for the evaluation of the economy and price stability, especially as for the compliance with the inflation criterion of Maastricht Treaty;

c) core inflation indices — prepared by the Central Bank of Poland for monetary policy purposes with the use of detailed data from the CSO consumer price survey.

All the above price indices are calculated as Laspeyres-type indicators. The only one computed annually according to the Paasche formula is for the deflation purposes of national accounts statistics.

12. Data publication

CPIs should be published and made available to users in the shortest possible time in accordance with international standards. The most frequently adopted deadline is about 2 weeks after the reference month. The Polish CSO observes this rule. In 2003 the dates of national CPI announcement were speeded up (on

average by 1-2 days) in order to adapt them to the schedule of data transmission to Eurostat and the release time of HICP data by the European Commission. National CPIs are announced as final data (except for January because of the change of the weighting system), while the HICP can be revised, i.e. it can be changed retrospectively following the introduction of methodological changes in the scope of index, etc.

Some statistical offices publish preliminary data on CPI, sometimes based on incomplete source data and after a few days they release final data. Since 2001 Eurostat has also been releasing preliminary data (the so-called "flash estimates") on inflation in the euro zone. Such data are published on the last day of the reference month or on the first or second day of the following month. They are prepared on the basis of the information from several large member states and on certain other items of price information. In the middle of the month following the reference month, the HICP and some other consumer price indices measured by Eurostat are released. If Eurostat retains these rules of data release in the following years, Poland should be prepared to provide the so-called early information for the purpose of "flash estimates".

13. Conclusions

Polish consumer price statistics, though evaluated by international institutions and main users as being of good quality (e.g. pre-accession missions of Eurostat's experts in 1999-2004; IMF mission in 2003), should be further developed. There is an obligation to implement efficiently the successive EU laws and improve survey practice on the basis of the experience of other member states, particularly those which joined the EU quite recently and had to revise their consumer price survey systems. Considering Polish membership of the EU, the CSO is obliged to treat compliance with Eurostat methodological requirements as a priority. However, methodological and organisational modifications of this survey for domestic needs should be implemented carefully, respecting the recommendations of the other international organisations (e.g. UN, OECD, ILO or IMF) to which Poland belongs.

In view of EU membership it seems reasonable to adopt the Eurostat HICP methodology also for the Polish national CPI, with a few reservations mentioned before. The economic aspect of the survey is of importance here. Having in mind careful management of resources it seems to be justified to apply in future some of the methodological principles, namely those relating to sample selection and price recording, as common for both surveys. The differences should concern some other aspects: system of weights, coverage, product scope, aggregation and data publication. These aspects of methodology should be formulated so as to meet the needs of the main users of data. Respecting the methodological principles of HICP at the stage of data collection also for CPI should not be in

contradiction to other methodological standards of the price survey, e.g. those recommended by ILO or UN ECE, which are more liberal.

International methodology is developing constantly, so further improvement of the consumer price statistics is likely to be needed, whatever the strategy of development. It is necessary to fulfil the duty of implementing new EU legislations, but also the revised ILO resolution on CPI as well as its updated CPI manual.

Improvements in price statistics methodology will also depend on further CPI harmonisation in the member countries. The concept of constructing a "family" of price indices based on HICP methodology would in practice mean adopting HICP as an official inflation measure and HICP methodology for the other consumer price indices produced by statistical offices for national needs. Possible measures to be taken would be aimed at increasing further the comparability of consumer price indices on the international scale. This would however require significant financial inputs for the modification of national CPI systems, including methodology, software and legislation requiring the Polish CSO to publish particular types of CPI.

Acknowledgement

The author is thankful to the referees for valuable comments and suggestions.

Views presented in the paper remain, however, the sole responsibility of the author.

REFERENCES

ALLEN R. G. D., (1975), *Index numbers in theory and practice*, Aldine Publishing Company, Chicago.

ASTIN J., (1999), *The European Union Harmonised Indices of Consumer Prices (HICP)*, Statistical Journal of the United Nations ECE, 16.

AT WHAT PRICE? Conceptualizing and measuring cost-of-living and price indexes, (2001) The National Academy of Science, National Academy Press, Washington DC.

BOSKIN M.J., DULBERGER E., GORDON R., GRILICHES Z., JORGENSON D., (1996), *Toward a More Accurate Measure of The Cost of Living*, Final Report to the US Senate Finance Committee, Government Printing Office, Washington, D.C.

CZAJKA B., NOWAK L., (1991), *Metodologia Obliczania Wskaznikow Cen Konsumpcyjnych* (Methodology of consumer price indices calculation), "Wiadomosci Statystyczne" vol. 2, Warsaw.

EUROSTAT (2001), *Compendium of HICP*, Reference documents, Commission of the European Communities, Eurostat.

GLUCHOWSKA A., (2002), *Metody hedoniczne w statystyce cen* (Hedonics methods in price statistics), "Wiadomosci Statystyczne" vol. 2, Warsaw.

GLUCHOWSKA A., TUREK D., (1993), *Ceny artykulow zywnosciowych* (Prices of food), "Wiadomosci Statystyczne" vol. 10, Warsaw.

GUS (1985), Zasady metodyczne statystyki cen detalicznych i kosztow utrzymania (Methodological rules for retail price statistics and cost of living statistics), Zeszyty metodyczne, vol. 54, Warsaw.

GUS (1990*), Statystyczne metody badania cen* (Statistical methods of a price survey), (1990), "BWS", vol. 38, Warsaw.

GUS (2002), *Zasady notowania cen detalicznych. wskazowki dla ankieterow Urzedow Statystycznych.* (Rules of retail price collection. Guidelines for price collectors in regional statistical offices), Warsaw.

HILL P., (1999), *Inflation, the Cost of Living and the Domain of a Consumer Price Index*, the Joint ECE/ILO Meeting on Consumer Prices, Geneva.

ILO (2001), *Consumer Price Indices,* Meeting of Experts on Labour Statistics, Geneva.

ILO (2003*),Consumer Price Indices*, ILO Manual, www.ilo.org.

ILO (2003), *Resolution Concerning Consumer Price Indices*, Seventeenth International Conference of Labour Statisticians, ILO, Geneva, www.ilo.org.

KORDOS J., (1988), *Badanie indeksow cen konsumpcyjnych w Polsce w swietle zalecen miedzynarodowych* (Consumer price survey in Poland in the light of international recommendations), "Wiadomosci Statystyczne" vol. 7, Warsaw.

KORDOS J., (1990), *O dokladnosci i precyzji indeksow cen konsumpcyjnych* (Accuracy and precision of consumer price indices), "Wiadomosci Statystyczne" vol. 5, Warsaw.

KORDOS J., CZAJKA B., (1987), *Consumer Price Indices in Poland*, Journal of the United Nations ECE, vol. 5.

KORDOS J., DUBLANKA I., (1982), *Metody badania cen towarow i uslug konsumpcyjnych w latach 1975-1982* (Methods for surveying consumer prices of goods and services in years of 1975—1982, "Wiadomosci Statystyczne" vol. 12, Warsaw.

KORDOS J., KOWALSKI G., (1983), *Program badan cen towarow i uslug konsumpcyjnych w warunkach silnej nierownowagi rynkowej* (A program of surveying consumer prices of goods and services in a highly unstable market), "Wiadomosci Statystyczne" vol. 1, Warsaw.

KUBICZEK F., (1990), *Statystyka cen detalicznych we Francji* (Retail price survey in France), "Wiadomosci Statystyczne" No 9, Warsaw.

LUSZNIEWICZ A., (1984), *Indeksy Kosztów Utrzymania, Teoria i Praktyka* (Cost of living indices. Theory and practice), Panstwowe Wydawnictwo Ekonomiczne, Warsaw.

NOWAK L., (1992), *Organizacja badania i metody gromadzenia informacji o poziomach cen towarow i uslug konsumpcyjnych* (Organisation of the survey and methods of price collection on consumer prices of goods and services), "Wiadomosci Statystyczne" vol. 2, Warsaw.

OKOLSKI M., TIMOFIEJUK I., (1981), *Statystyka ekonomiczna. Elementy teorii* (Economic statistics. Elements of Theory), Panstwowe Wydawnictwo Ekonomiczne, Warsaw.

ROMANIUK K., (1976), *Indeksy dynamiki zjawisk ekonomicznych* (Indices of dynamics of economic phenomena), Panstwowe Wydawnictwo Ekonomiczne, Warsaw.

ROSC Report by IMF, (2003), www.imf.org.

SMILOWSKA T., (2001), *Proba uwzglednienia jakosci w badaniu cen* (An attempt at introducing quality change in the price survey), "Wiadomosci Statystyczne", vol. 5, Warsaw.

SMILOWSKA T., CYNKIER A., (2002), *Metody uwzglednienia jakosci w badaniu cen* (Methods of introducing quality change in the price survey), "Wiadomosci Statystyczne", vol. 8, Warsaw.

TOCZYNSKI T., (1987), *Ceny urzedowe, regulowane i umowne. Tendencje zmian w latach 1985—1986* (Prices: administrated, regulated and contracted. Trends in years 1985—1986), "Wiadomosci Statystyczne", vol. 10, Warsaw.

TOCZYNSKI T., (1989), *Badanie cen detalicznych. Rozwiazania obecne i propozycje modyfikacji.* (Retail price survey. Current solutions and proposals for modifications), "Wiadomosci Statystyczne", vol. 10, Warsaw.

TUREK D., (1992), *Metodologia obliczania wskaznikow cen konsumpcyjnych i niekonsump-cyjnych* (Methodology of calculations of consumer and non-consumer price indices), "Wiadomosci Statystyczne", vol. 2, Warsaw.

TUREK D., (1995), P*rzeglad praktycznych rozwiazan stosowanych w badaniu cen konsumpcyjnych w Polsce* (A review of the practices introduced in the consumer price survey in Poland), "BWS", vol. 44, Warsaw.

TUREK D., GLUCHOWSKA A., (1995), *Jak Sie Liczy Ceny?* (How prices are calculated?), Statystyka Polski — dodatek do dziennika "Rzeczpospolita" No 1(71), Poland.

TURVEY R., (1989), et al. *Consumer Price Indices,* An ILO Manual, Geneva.

UNITED NATIONS (2000), *Classification of Individual Consumption According to Purpose*, United Nations, New York.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1151—1163

THE INFORMATION CONTENT OF ANALYSTS' RECOMMENDATIONS: THE EMPIRICAL EVIDENCE FROM THE WARSAW STOCK EXCHANGE

Henryk Gurgul[1] and **Paweł Majdosz**[2]

ABSTRACT

In this article we investigate the impact of analysts' recommendations on stock prices using data from the Polish stock market. To improve market transparency, our interest in this subject is limited to those companies listed on the Warsaw Stock Exchange (WSE) whose shares are the most liquid. Applying event study methodology, we define abnormal return as a difference between actual return of a given firm and expected return generated by the market model. The results show that *buy* as well as *hold* recommendations have no informational content for the investors operating on the WSE. But contrary to this, the shareholders of firms from panel *Sell* experienced a mean abnormal return of –1.01% on day –1. Robustness of our findings is confirmed by using bootstrap technique. This study is the first of this kind for the Polish stock market, and to our knowledge also the first for the emerging markets from Central and Eastern Europe.

***Key words*:** Event study, Analysts' recommendation, Market model, Bootstrap.

1. Introduction

In his contribution Fama (1976) formulated the well–known Efficient Market Hypothesis (EMH). He described an efficient market as that having prices that "fully reflect" available information. In the context of this hypothesis many researchers started to estimate the information content of different events such as earnings announcements, dividend announcements, buy back, stock split, and others (see e.g. Mestel et al., 2003).

[1] [2] Department of Applied Mathematics, University of Science and Technology, Al. Mickiewicza 30, 30-059 Kraków, Poland; E–mail: h.gurgul@neostrada.pl.

Information is costly to process. Brokerage firms spend hundreds of millions of dollars annually analyzing stocks and trying to persuade investors that certain stocks are more or less attractive than others. Investors should be willing to pay for brokerage investment advice only if the expected benefit is at least as great as the cost of the advice.

With access to information, the increased number of analysts' recommendation announcements can be observed, especially over a few recent years. The large number of publications supplies factual news and opinions concerning unique investment opportunities to investors. Not all analysts' recommendations are, however, based on serious research. Sometimes the goal of such a recommendation is to manipulate of stock prices for private advantages. Market participants are likely not to distinguish between reliable and manipulated analysts' recommendations. This is why in some cases market reaction to analysts' recommendation announcements may be rather confusing.

Many countries try to hold the neutrality of research being the basis of recommendation announcements. There exist many institutional regulations of the *Chinese Wall* type such as compliance department or black out periods whose target is to preserve the neutrality of such research. On the other hand, serious effort has been made by researchers to throw light on the issue of the possible impact of professional analysts on stock market.

The quality of analysts' earnings forecasts is one of the three research directions (see e.g. Womack, 1996). Womack examines a comprehensive set of recommendations from the fourteen U.S. brokerage firms. He tries to evaluate the ability of analysts to predict or influence stock prices. The second very important topic is the quality of professional security analysis of investment funds. According to Malkiel (1995), the active funds demand management fee and therefore the higher risk–adjusted returns can be expected. The impact of analysts' recommendation announcements on stock prices is the third research direction. The rating change announcements released, for example, by Moody or Standard&Poor are the well–known exemplification of such an effect (see e.g. Hand et al., 1992). The effect of investments of *Value Line* on common stock prices is presented in Stickel (1985).

Loh et al. (1990) examine whether the publications of Business Week's *Inside Wall Street* column (IWS) affect stock returns of companies mentioned in it. Assuming that market is efficient in the semi–strong form, if this column uses information that are commonly available for market participants, no one should be able to earn abnormal profits on the basis of prior knowledge about the column content. An examination of stock abnormal returns reveals that, except for *Buy* sub–sample, there is no evidence of market response to the IWS publications. Two possible implications of it have been given. Firstly, it is possible that the IWS column is constructed on the basis of commonly available information. Secondly, investors seems to be reluctant to take actions when a given firm is presented in a negative or neutral light. The more recent study which deals with

the analysts' impact on investors was performed by Gerke and Oerke (1998) using German data.

This study is aimed at testing the informational content of analysts' recommendation announcements. To accomplish our intention, we use event study methodology based on the set of analysts' recommendation announcements which were released on the WSE during the period 03/07/1995–30/09/2003. In connection with the main goal of our contribution, we formulate the conjectures which are checked under this study as follows.

1. *Buy* recommendations are considered by market participants as a positive signal. Therefore, we hypothesize that when it is recommended buying shares of a given firm, its stock prices increase significantly.
2. After *hold* recommendation announcements, investors are prone to do nothing. Hence, the zero effect on stock prices should be expected.
3. *Sell* recommendations are likely to be a negative signal for investors. Therefore, it is hypothesized that sell recommendation announcements are accompanied by decreases in stock prices.

The Polish stock market is often branded as an insider market where some investors receive new information earlier than others. There also arises a question about the extent to which the Warsaw Stock Exchange (WSE) is manipulated and dominated by rumors and inside information.

Our results show that *buy* as well as *hold* recommendations have no informational content for the investors operating on the WSE. But contrary to this, the shareholders of firms from cluster *Sell* experienced an abnormal return of —1.01% on day —1. This finding can be interpreted as evidence of leakage of information prior to the release of analysts' recommendation announcements.

The paper is organized as follows. Section 2 demonstrates our database. In Section 3 we describe in detail the methodology applied within the frame of event study. Section 4 contains empirical results including parametric test results as well as bootstrap inferences. Section 5 concludes.

2. Data and sample selection

To be included in our sample, a given firm has to meet some selection requirements. Firstly, it should be reliable and large enough to be quoted on the primary market of the WSE. There is of course no one–to–one relation between firm liquidity (or size) and manipulation of analysts' recommendations, but there is less likelihood of such a manipulation for large firms, since they are more extensively covered by many independent analysts. Therefore, secondly, a firm which is a candidate for the sample should be included in the WIG20. The WIG20 comprises twenty stocks which are the most liquid in Poland. Of course, there is no need for the sampled firms to be quoted in the WIG20 in the whole analysed period. Instead, we demand from it to be included in the WIG20 over at least

three months when analysts' recommendation announcements took place. These requirements fulfill 19 companies listed in Table 1.

We define the announcement (event) date as the occasion of the very first recommendation of professional institutions such as banks and broker offices that can be identified in our database. Sources for these data include: the register of analysts' recommendation announcements which can be accessed for free by Internet at http://www.parkiet.com.pl, and other daily as well as periodical professional publications. It is very important to make clear that the very first recommendation means the first time when it is recommended buying, holding or selling shares of a given firm. Therefore, all the cases where bank or broker office release recommendation announcement which is in line with the previous one, have been ignored by us under this study.

From 3rd July 1995 through 30th September 2003, the number of analysts' recommendation announcements for sampled firms totals 2,880. After excluding recommendations which were in line with the previous one, and that affected by confounding events such as earnings or dividend announcements, 139 announcements remain. Selected analysts' recommendation announcements were categorized according to the direction of changes in position of investors which is recommended by analysts into *buy*, *hold* and *sell*. Table 1 presents the number of analysts' recommendation announcements which were released on the WSE for each year in the analysed period as well as the number of selected events.

Table 1. Analysts' recommendation announcements on the WSE in the period from July 1995 to September 2003*

Year	Total number of recommendation announcements	Number of events		
		Buy	Hold	Sell
1995	13	0	0	0
1996	27	1	1	0
1997	66	2	0	3
1998	108	6	4	0
1999	660	6	11	1
2000	990	7	12	3
2001	587	14	8	2
2002	296	12	14	9
2003	132	8	9	6
Σ	**2880**	**56**	**59**	**24**

* Only for firms included in our sample.

Daily closing prices for firms included in our sample are derived from the PARKIET database. In a few cases a single missing piece of data was filled in (as a mean of its direct neighbours).

3. Methodology

At the beginning of our study different computational schemes of stock price changes (i.e. returns) were considered. Two alternative formulas are defined below:

$$R_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}} = r(t-1,t) \tag{1}$$

$$R_{i,t} = \log\left(P_{i,t}\right) - \log\left(P_{i,t-1}\right) = \log\left[1 + r(t-1,t)\right] = \tilde{r}(t-1,t) \tag{2}$$

where $P_{i,t}$ stands for the stock price of company i on date t, $P_{i,t-1}$ stands for the stock price of company i on date t–1, and *log* denotes the natural logarithm. As we see $\tilde{r}\left(t-1,t\right)$ is an approximation of the theoretical definition of returns given by (1). Taking the second-order Maclaurin expansion of (2) , we can state that this formula undervalues the true return (1). In spite of this we decided to compute stock returns by using equation (2), because the use of a natural logarithm improves some statistical properties of financial time series distribution. Furthermore, since the logarithmic transformation belongs to the Box–Cox transformation, it can stabilize the variance.

To apply event study methodology, returns are computed over 35 trading days, i.e. from t = –32 to t = +2 (t = 0 is defined as event day, i.e. the analysts' recommendation announcement day). We are interested in security price behavior in the so–called event window, which comprises 5 trading days, i.e. the two days prior to the announcement release, the event day, and the two following days.

In the next step, we calculate the abnormal returns ($AR_{i,t}$) which are defined as a difference between observed return and expected return for day t. The expected value of stock returns is generated based on the market model in which the stock returns of a given firm depend on market portfolio returns. This can be expressed as:

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{i,t} \tag{3}$$

where $R_{i,t}$ denotes the return of stock i on day t, and $R_{m,t}$ stands for the return of market portfolio. To proxy market portfolio returns, returns of the market–capitalization weighted stock index called WIG are employed.

The equation (3) was estimated on the basis of data from the so–called pre–event window (i.e. for t = –32,…,–3) by using the OLS–method. The advanced models with heteroscedastic error term require at least 50 observations to be successfully estimated. To guarantee this length of the pre–event window, we would have to reduce drastically the number of events in each clusters. It would be impossible to carry out any statistical analysis of the abnormal returns, because the number of events across clusters would be insufficient to do this. In addition, the standard market model given by (3) is commonly used by many researchers to

forecast individual stock returns. On the other hand, it is a widely–documented fact that the crucial assumptions about probabilistic properties of model error term are invalid in the majority of cases where this model is practically employed. To avoid a possible distort of our inferences by heteroscedasticity or serial correlation in residuals, or both these effects, we conduct the post–estimation analysis.

At the beginning, we formulate the null hypothesis that the OLS residuals are normally distributed. By means of the Jarque–Bera test, we find twenty three cases where the distribution of the OLS–residuals is significantly different from normal.

After controlling for normality, we continue testing by using the Ljung–Box test for model misspecification due to autocorrelation which is based on the Q statistic defined below:

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \varepsilon_{i,t} \tag{4}$$

where I stands for the number of observations, L stands for the number of autocorrelations (chosen $L = 15$), r_j^2 denotes a squared autocorrelation of the residuals at lag j. Under the null hypothesis statistic (4) is asymptotically χ^2 distributed. Only in fourteen cases out of a total number of one hundred thirty nine, test statistic Q is significant.

Finally, we checked whether the variance of the residuals can be assumed to be constant over time. To confirm or reject the null hypothesis that the OLS–residuals do not exhibit the heteroscedastic effect, the White test is performed. We find thirty two cases where the value of test statistic contradicts the null.

Concluding the post–estimation analysis, we find that about 25 percent of cases do not meet our requirements. However, the OLS–estimator still remains consistent even though some assumptions are invalid, but its efficiency is low. Overall, the use of market model under this study seems to be justifiable.

Testing the informational content of analysts' recommendation announcements is conducted separately in the three clusters, namely: *Buy*, *Hold* and *Sell*. For all days within the event window (i.e. for $t = -2,\ldots,+2$), we check the null hypothesis about zero average abnormal returns against the alternative that average abnormal returns are statistically different from zero by using a test based on the t–statistic given as:

$$t = \frac{\frac{1}{N^k}\sum_{i=1}^{N^k} AR_{i,t}}{\hat{\sigma}_{AR}}, \tag{5}$$

where $AR_{i,t}$ denotes abnormal return of ith firm for day t, N^k stands for the number of events in cluster k, and $\hat{\sigma}_{AR}$ is defined as:

$$\hat{\sigma}_{AR} = \frac{1}{N^k}\sqrt{\frac{1}{29}\sum_{t=-32}^{-3}\left(\sum_{i=1}^{N^k} AR_{i,t} - \sum_{t=-32}^{-3}\sum_{i=1}^{N^k} AR_{i,t}\right)^2}. \qquad (6)$$

Note that the standard deviation (6) is computed within the pre–event window to avoid an increase in variance of abnormal returns when the event occurs. Under the null, the statistic (5) is t distributed with $N^k - 1$ degrees of freedom.

4. Empirical results

Excess return reaction

Our test results for each day within the event window in the three clusters are summarized in Table 2. This table contains the average abnormal returns, values of statistic (5) as well as the corresponding p–values. At the beginning, we check whether the average abnormal returns are not serially correlated and distributed as normal by using the above–mentioned Ljung–Box test and Jarque–Bera test. In both cases there is no reason to reject the null hypothesis at a 5% level of significance.

From the figures in Table 2, one can find that test statistics are insignificant in the first two clusters (Buy and Hold), which means that investors are reluctant to change their positions in stocks of a given firm when this firm is presented by analysts in a positive or neutral light. On the other hand, the shareholders of companies from cluster Sell realize an average abnormal return of –1.01% on one day prior to the announcement release. This finding indicates that market participants receive the news prior to the announcement release and therefore analysts' recommendation arrival does not cause any visible stock price movements on the event day.

Table 2. Testing of daily abnormal returns in the three clusters

	Event day *t*				
	–2	–1	0	+1	+2
			Buy		
Average abnormal return (%)	0.304	-0.383	0.065	0.195	-0.103
t–Student statistic	0.949	-1.194	0.203	0.609	-0.320
p–value	0.347	0.238	0.840	0.545	0.750
			Hold		
Average abnormal return (%)	0.343	0.290	-0.109	-0.127	-0.015
t–Student statistic	1.141	0.964	-0.363	-0.421	-0.050
p–value	0.259	0.339	0.718	0.675	0.960
			Sell		
Average abnormal return (%)	0.288	-1.009*	-0.128	-0.383	0.224
t–Student statistic	0.755	-2.641	-0.336	-1.002	0.586
p–value	0.458	0.015	0.740	0.327	0.564

* Significant at 5% level

Sample size in the respective clusters are equal to 56, 59 and 24.

Our results (for the emerging stock market like the Polish one) are not fully in line with those of Loh et al. (1990) and Gerke and Oerke (1998) which are presented for the developed the U.S. and the German stock markets, respectively. In the work of Loh et al., *neutral* and *sell* recommendations are generally accompanied by the zero impact on stock prices while *buy* recommendations cause significant abnormal returns around the event day. On the other hand, Gerke and Oerke argue that *neutral* and *sell* recommendations induce significant abnormal earnings for shareholders. One can notice that our findings are placed somewhere between those of Loh et al. and those of Gerke and Oerke. In addition, our findings corroborate the widely–documented fact that analysts' recommendation announcements influence stock prices weaker than earnings or dividends announcements. The reason is that investors are likely to have problems with the correct interpretation of such announcements. It seems to be particularly true when analysts recommend that invertors should buy shares of a given firm, because in such cases there is great likelihood that the advice is motivated by private advantages.

Bootstrap test

The validity of our *t*–statistics is supported by application of a bootstrap test. To obtain the bootstrap analogue of (5), we need to generate data from the market

model under the null of the zero impact of analysts' recommendation announcements on stock prices. First, the OLS–residuals are resampled (with replacement) for the pre–event window, and then the bootstrap sample $\left\{\varepsilon_{i,t}^{*}\right\}$ is completed by drawing (also with replacement) abnormal returns for each day within the event window from the OLS–residuals.

Let $R_{i,t}^{*} = \hat{\alpha}_i + \hat{\beta}_i R_{m,t} + \varepsilon_{i,t}^{*}$, for $t = -32,\ldots,+2$. In the next step, applying the OLS–method, we estimate the parameters of model (3) in which $R_{i,t}^{*}$ is used instead of $R_{i,t}$. Finally, the bootstrap analogue (t^b) of test statistic (5) is calculated for each day within the event window. This procedure is repeated a large number of times (chosen number of iterations $N^b = 10{,}000$).

The bootstrap critical significance level (p–value^b) is calculated as:

$$p - value^b = \frac{\operatorname{card}\left\{t^b \in \left\{t_1^b,\ldots,t_{N^b}^b\right\} : \left|t^b\right| \geq \left|t\right|\right\}}{N^b}. \tag{7}$$

The bootstrap test results are reported in Table 3. Note that results are organized in three separate panels, for *buy*, *hold* and *sell* recommendations.

One can find that the bootstrap p–value is less than a 5% level on one day prior to the analysts' recommendation announcement, in which it is recommended selling shares of a given firm. In the remaining two clusters there is no significant event effect. This finding fully corroborates the statistical inference on the basis of test statistic (5).

Table 3. Summary statistics for t^b in the three clusters

	Event day *t*				
	–2	–1	0	+1	+2
			Buy		
Min	-5.852	-4.513	-5.460	-6.559	-4.842
1. Quartil	-0.771	-0.752	-0.746	-0.772	-0.756
Median	-0.038	-0.037	-0.042	-0.050	-0.040
2. Quartil	0.682	0.691	0.665	0.665	0.679
Max	4.406	4.226	4.521	5.318	4.331
p–value[b]	**0.376**	**0.264**	**0.844**	**0.564**	**0.760**
			Hold		
Min	-5.057	-4.588	-4.958	-5.977	-4.649
1. Quartil	-0.791	-0.751	-0.799	-0.762	-0.782
Median	-0.083	-0.033	-0.079	-0.062	-0.037
2. Quartil	0.666	0.664	0.660	0.655	0.663
Max	4.791	4.774	4.667	4.289	4.436
p–value[b]	**0.289**	**0.360**	**0.744**	**0.686**	**0.958**
			Sell		
Min	-5.316	-4.526	-5.113	-5.518	-4.436
1. Quartil	-0.774	-0.793	-0.798	-0.796	-0.822
Median	-0.046	-0.094	-0.107	-0.084	-0.102
2. Quartil	0.646	0.615	0.612	0.616	0.606
Max	6.426	4.296	5.089	5.595	4.674
p–value[b]	**0.471**	**0.020**	**0.748**	**0.342**	**0.581**

5. Conclusions

Our study is the first one to analyse the informational content of analysts' recommendation announcements for the Polish stock market. To improve market transparency, our interest in this subject is limited to the companies whose shares are quoted in the WIG20. The event study used here provides evidence which can be summarized as follows.

1. There is no event effect when analysts recommend that investors should buy or hold shares of a given firm.
2. If analysts recommend that investors should sell shares of a given firm, negative market response to this news is noticeable on one day prior to the announcement release. This supports the view that investors are rather reluctant to change their position in stocks on the basis of analysts' recommendations. One of possible explanations for this is that analysts'

recommendations are based on information which is commonly available for market participants. Knowing information prior to the analysts' recommendation announcement release, investors have no reason to correct their expectations when it comes in. Another explanation focuses our attention on problem of reliability of such a recommendation. It is very likely that market participants do not distinguish between reliable and manipulated analysts' recommendations. Therefore, market reaction to analysts' recommendation announcements may be unclear. The evidence that stock prices react significantly to analysts' recommendations on day –1 in *Sell* cluster leads to the question whether an analyst becomes an insider due to his knowledge expressed by him on the announcement day.

Our study is limited by the low number of events included in the sample, especially, in *Sell* cluster. Therefore, it seems to be needed to confirm this part of presented results by future research.

Appendix

List of companies included in the sample and the number of events

Company /ISIN/	Analysts' recommendation		
	Buy	Hold	Sell
AGORA /PLAGORA00067/	5	1	2
BUDIMEX /PLBUDMX00013/	7	4	1
BANK BPH /PLBPH0000019/	3	1	2
BRE /PLBRE0000012/	3	3	1
BZ WBK /PLBZ00000044/	2	3	-
COMARCH /PLCOMAR00012/	1	2	4
COMPUTERLAND /PLCMPLD00016/	4	2	1
DĘBICA /PLDEBCA00016/	3	3	1
ŚWIECIE /PLCELZA00018/	6	3	1
KGHM /PLKGHM000017/	-	4	1
KĘTY /PLKETY000011/	4	8	2
BIG BG /PLBIG0000016/	-	5	2
ORBIS /PLORBIS00014/	3	2	1
BANK PEKAO /PLPEKAO00016/	1	1	-
POLSKA GRUPA FARMACEUTYCZNA /PLMEDCS00015/	6	6	-
PROKOM SOFTWARE /PLPROKM00013/	3	4	1
PKN ORLEN /PLPKN0000018/	1	4	1
SOFTBANK /PLSOFTB00016/	2	2	2
TELEKOMUNIKACJA POLSKA /PLTLKPL00017/	2	1	1

REFERENCES

FAMA, E.F. (1976), *Foundations of Finance* (New York, Basic Books).

GERKE, W., OERKE, M. (1998), Marktbeeinflussung durch Analystenempfehlungen, *Zeitschrift fuer Betriebswirtschaft*, 2, pp. 187—200.

HAND, J.R.M., HOLTHAUSEN, R.W., LEFTWICH, R. (1992), The Effect of Bond Rating Announcement on Bond and Stock Prices, *Journal of Finance*, vol. 47, pp. 732—752.

LOH, Ch., PERUMPRAL, S., KEOWN, A.J. (1990), The Impact of the "Inside Wall Street" Publications on Stock Returns and Trading Volume, *American Business Review*, pp. 41—49.

MALKIEL, B.G. (1995), *Returns* from Investing in Equity Mutual Funds 1971-1991, *Journal of Finance*, vol. 50, pp. 549—572.

MESTEL R., GURGUL H., SCHLEICHER, C. (2003), *Capital market efficiency - an empirical analysis of the dividend announcement effect for the Austrian stock market,* Operations Research 2002, Springer-Verlag, Berlin, Heidelberg, New York, Hong Kong, London, Milan, Paris, Tokyo, pp. 315—320.

STICKEL, R.J. (1985), The Effect of Value Line Investment Survey Rank Changes on Common Stock Prices*, Journal of Financial Economics*, vol.14, pp. 121—144.

WOMACK, K.L. (1996), Do Brokerage Analysts' Recommendations Have Investment Value?, *Journal of Finance*, vol. 51, pp. 137—167.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1165—1174

CARE NEEDS AND CLINICAL CHARACTERISTICS OF IN-PATIENTS IN SOUTHERN ESTONIA[1]

Gunnar Ljunggren[2], Kai Saks[3], Ene-Margit Tiit[4], Björn Osberg[5], Toomas Kivastik[6]

ABSTRACT

The aim of our survey was to determine the structure of comprehensive needs of patients admitted to acute care hospitals. The sample consisted of all in-patients (n=1535) in all 15 acute care hospitals in the region of south-east Estonia. The patient classification system Resource Utilization Groups (RUG-II) was used to estimate different clinical needs. The occupancy rate of beds was 89%. Of all patients 36% were ready for discharge. These patients had less needs by RUG-II categories in clinical complexity and special care, but more in basic nursing. Approximately one quarter of hospitalized patients did need some type of care after discharge from acute care, whereas more than half of them needed nursing or community care services. In conclusion, our data indicate that there is still space for optimizing resource use in south-east Estonia by developing post-acute and long-term care services, based on a comprehensive assessment.

Key words: occupancy rate of beds; length of stay; RUG-categorization.

1. Introduction

The health care burden is increasing in many countries, mainly because of the population ageing. Unfortunately there is yet no empirical support for the compression theory suggesting that medical progress will reduce the duration of morbidity during life more significantly than increasing morbidity by extending life, as among others demonstrated Yip and Law in 2002.

[1] This work was partly supported by ESF Grant 5521.

[2] Centre for Gerontology and Health Economics, Karolinska Institutet, Stockholm, Sweden.
[3] Department of Internal Medicine, University of Tartu, Tartu, Estonia.
[4] Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia.
[5] Scandinavian Care Consultants Services, Ltd., Stockholm, Sweden.
[6] Institute of Pharmacology, University of Tartu, Estonia.

To optimize resource use within the health care sector, there is a trend to reduce the average length of stay of patients in acute care hospitals all over the world and provide services in other institutions or at home, see Clarke and Rosen (2001) and European observatory on health care systems, (2000). But to achieve good results and cost-effectiveness, it is necessary to secure that a diverse and flexible system of post-acute and nursing care with integrated health and social care services is implemented, particularly for the frail elderly, as suggested Hansen et al (1992), Melin and Bygren, (1992) and also Nikolaus et al (1999).

In Estonia, which is a country passing through economic and social transition, the reform of health care started ten years ago with rearrangement of primary care and acute care. In the year 2000 there were 7258 acute care beds in Estonia (approximately 5 beds per 1000 inhabitants). According to the plans there will be 3100 acute care beds in 2015 (2 beds per 1000 inhabitants), see Estonian nursing care master plan 2015 (2001). Development of long-term care was initiated as late as in 2001. For planning of long-term care services, data on the medical, functional, social and psychological needs of patients must be taken into consideration.

The aim of our survey was to determine the structure of physical, cognitive, social and medical needs of patients admitted to acute care hospitals. In addition, the aim was to assess what kind of care should be offered to patients after discharge.

2. Methods

The study was performed in south-east Estonia, one out of four medical regions in the country. The region includes 350 000 inhabitants (26% of the whole population of Estonia) and 15 hospitals for acute care. The leading medical centre for the region is the University Hospital in Tartu. All hospitals were included in the survey with the only exception of children's hospital.

Data were collected for every patient admitted at least one day prior to the study including those who were expected to be discharged on the day of the survey. Personal data (gender, age, date of admission, living site prior to admission) were recorded by the chief nurses of the departments. Medical status (main diagnosis from a list of ten broad groups, and if the patient was ready for discharge), functional status (activities of daily living — eating, transferring, continence) and needs for services after discharge were assessed by physicians in cooperation with chief nurses. The patient classification system Resource Utilization Groups (RUG-II) was used to estimate different clinical needs by Schneider et al (1988).

Chief nurses and medical directors of all hospitals were carefully instructed during a seminar before the study, written instructions were distributed and phone consultation offered. The study was performed during one day in every department, in May 2001.

3. Statistical analysis

The patients included in the study were divided in two subgroups in each type of department, on the basis of being ready for discharge or not. The difference between subgroups was analysed using *t*-test for continuous variables and χ^2-test for categorical variables.

Differences between the types of beds were assessed by one-factorial dispersal analysis. P-values <0.05 were considered significant. All tests were performed using SAS-® statistical system.

4. Results

During the days of the survey, there were 1535 patients in 1722 beds in acute care hospitals for surgery, internal medicine, gynaecology and psychiatry in the region (occupancy rate 89%). In surgery beds 21.3% of patients had cancer, 8.2% musculosceletal disorders, 6.4% neurological diseases, 2.3% hip fractures, 1.0% stroke. In internal medicine beds 16% of patients had disorders of heart and circulation, 9.7% stroke, 7.5% other neurological disorders, 5.1% cancer, 5.0% musculosceletal diseases, 4.0% diabetes, 3.1% psychiatric disorders, 2.9% dementia. Of all gynaecological patients 14.6% had cancer. Among patients in psychiatric beds 4.2% had dementia and 87.2% other psychiatric disorders. The difference between two subgroups — ready for discharge and not ready for discharge — has been tested in each type of beds.

Characteristics of the patients are presented in Table 1. Patients in gynaecology and psychiatry were younger than in surgical and internal medicine departments (P<0.05), where approximately half of patients were older than 65 years. The length of stay from admission to assessment was extremely long in psychiatry, 568 days on average. Patients in internal medicine departments had longer stays than in surgery (10 vs. 6 days; P<0.05). In gynaecology beds the length of stay was the shortest (4 days). Patients admitted to internal medicine or psychiatric departments were more often living alone than patients in gynaecology and surgery departments.

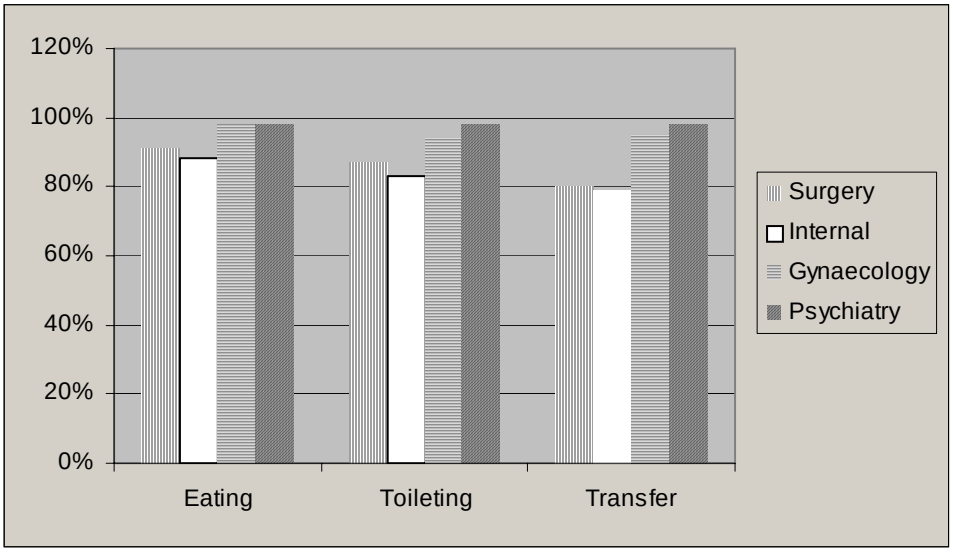
Of all examined patients, 36% were ready for discharge with no big differences between surgery and internal medicine, whereas in psychiatry most of the patients were too sick to be discharged and in gynaecology most of the patients were ready for discharge.

Table 1. Basic characteristics of the patients in acute care hospitals in Southern Estonia. Difference between two subgroups — *ready for discharge* and *not ready for discharge* — in each type of beds: * P< 0.05; ** P>0.01, *** P<0.001

Type of beds (number of beds)	Age Mean ±SD	Patients 65+ %	Proportion of women %	Time from admission to assessment, days Mean±SD	Time from admission to discharge (when estimation possible), days Mean±SD (estimated persons)	Living alone %	Too sick to be discharged %
Surgery (512)	58±19	42	48	6±8	10±10 (332)	16.0	60
Ready for disch (203)	59±18	44	50	6±7	7±7 (178)***	17.2	
Remaining (309)	57±19	40	47	7±9	13±12 (154)	15.2	
Internal (884)	64±17	56	59	10±18	15± 36 (675)	28.8	60
Ready for disch (358)	64±16	56	59	12±23*	11±11 (315)***	29.9	
Remaining (526)	63±17	56	59	9±13	19±48 (360)	28.1	
Gynaecol (48)	37 ±17	8	100.0	4±4	4±3 (35)	4.2	27
Ready for disch (35)	35±15	6	100.0	3±3	4±4 (27)	2.9	
Remaining (13)	41±20	15	100.0	5±6	4±3 (8)	7.7	
Psychiatry (289)	48 ±16	15	48	568±1965	24±18 (96)	40.5	88
Ready for disch (34)	50±16	18	38	1081±3006	22±35 (21)	41.2	
Remaining (255)	48±16	15	50	488±1746	25±9 (75)	40.4	
All (1733)	59 ±18	43	55	91±780	14±29 (1138)	27	64
Ready for disch (630)	60±18*	47*	57	66±722	10±12 (541)***	26	
Remaining (1103)	58±18	41	54	107±814	18±38 (597)	28	

Ability of patients to fulfil activities of daily living in eating, toileting and transfer is presented in Figure 1.

Figure 1. Percent of patients' independent in three Activities of Daily Living



The measures showed a low dependency rate regarding physical functions: 90% were independent in eating, 86% in transfer, and 84% in toileting. Patients in surgery and internal medicine beds had higher dependency than in gynaecology and psychiatry. In surgery, internal medicine and gynaecology the independency rate was significantly higher in those who were ready for discharge than in those that were not in eating 4–8% (P=0.005 in surgery; P=0.002 in internal medicine; P=0.073 in gynaecology), toileting 5–12% (P=0.031; P=0.008; P=0.017) and transfer 6–15% (P=0.029; P<0.001; P=0.320). In psychiatry there were no significant differences between these two groups of patients.

The results of the assessments of type and intensity of care by RUG-II categories are shown in Table 2. In surgery, internal medicine and gynaecology clinical complexity (need for medical attention and nursing care) was the dominating need whereas in psychiatry behavioural problems and basic nursing were more prevalent. Those who were considered ready for discharge in surgery and internal medicine had less needs in clinical complexity and special care, but more in basic nursing.

Table 2. Classification of patients by need and intensity of care (RUG-II categories, in percentages). Difference between two subgroups — *ready for discharge* and *not ready for discharge* — in each type of beds: * P< 0.05; ** P>0.01, *** P<0.001.

Type of beds (number of beds)	Rehab	Special care	Clinically complex	Beha-viour	Physical
Surgery (512)	13.9	5.9	48.0	0.4	31.8
Ready for discharge (203)	16.3	1.5***	32.0***	1.0	49.3***
Remaining (309)	12.3	8.7	58.6	0.0	20.4
Internal (884)	24.8	5.5	48.9	3.2	17.6
Ready for discharge (358)	28.5*	3.9	40.5***	4.5	22.6***
Remaining (526)	22.2	6.6	54.6	2.3	14.3
Gynaecology (48)	0.0	16.7	83.3	0.0	0.0
Ready for discharge (35)	0.0	20.0	80.0	0.0	0.0
Remaining (13)	0.0	7.7	92.3	0.0	0.0
Psychiatry (289)	9.0	12.1	40.5	18.3	20.1
Ready for discharge (34)	11.8	0.0	41.2	23.5	23.5
Remaining (255)	8.6	13.7*	40.4	17.6	19.6
All (1733)	18.2	7.0	48.2	4.8	21.8
Ready for discharge (630)	22.0**	3.8***	40.0***	4.1	30.0***
Remaining (1103)	16.0	8.8	52.8	5.1	17.0

The structure of post-discharge needs for services of the patients is presented in Table 3. Our data indicate that one quarter of hospitalized patients did need some type of medical or social care services after discharge. Approximately 10% of patients in surgery and internal medicine needed further health care services in another acute or rehabilitative/geriatric department and 15% of internal medicine and psychiatric patients in nursing facilities or community care services.

Table 3. Needs for services after discharge (in percentages). Difference between two subgroups — *ready for discharge* and *not ready for discharge* — in each type of beds: * P< 0.05; ** P>0.01, *** P<0.001.

Type of beds	Other hospital	Rehab/ geriatrics	Nursing home	Commu-nity care	Home	Other
Surgery (512)	6.4	5.5	1.8	3.7	70.7	2.0
Ready for discharge (203)	4.9	4.9	2.0	5.4	79.3***	2.0
Remaining (309)	7.4	5.8	1.6	2.6	65.0	1.9
Internal (884)	7.2	2.7	6.0	8.9	69.7	1.4
Ready for discharge (358)	5.9	3.1	3.9*	5.9**	77.4***	0.0**
Remaining (256)	8.2	2.5	7.4	11.0	64.4	2.3
Gynaecology (48)	0.0	0.0	0.0	2.1	97.9	0.0
Ready for discharge (35)	0.0	0.0	0.0	2.9	97.1	0.0
Remaining (13)	0.0	0.0	0.0	0.0	100.0	0.0
Psychiatry (289)	2.1	0.7	5.9	9.0	40.5	0.7
Ready for discharge (34)	11.8***	5.9***	35.3***	0.0	44.1	0.0
Remaining (255)	0.8	0.0	2.0	10.2	40.0	0.8
All (1733)	5.9	3.1	4.6	7.2	65.9	1.4
Ready for discharge (630)	5.6	3.7	4.8	5.2*	77.3***	0.6*
Remaining (1103)	6.2	2.8	4.4	8.3	59.4	1.8

5. Discussion

According to a recent international report European observatory on health care systems (2000), the occupancy rate of acute care beds varies to a large extent in European countries, ranging from 27% in Georgia to 94% in Israel, being most frequently between 70 and 90%. Our data revealed a quite high occupancy rate (89%) in the south-east region of Estonia. The length of stay exceeded the same index of the majority of European countries, see Madsen et al, (2002). At the same time only 64% of patients were considered to be too sick for discharge. Inappropriate hospital stay could probably never be reduced to zero but in south-east Estonia it is almost two-fold higher when compared with other countries, see Vaughan and Withers (2002) and Panis et al (2002). At the same time patients in Estonian hospitals had relatively good functional status. These data indicate that there is obviously still space for reducing length of stay and free more beds in acute care hospitals. Coughlan and O'Neill (2001) declared by Irish data and Onen et al (2001) using data from France that sixteen percent of the patients who needed medical, nursing and social care after discharge could be discharged probably much earlier if there existed flexible system of geriatric post-acute and nursing care. Since the patients with rehabilitation needs and those with medical complexities over-weighted the patients with only physical dependencies in surgery and internal medicine departments, where half of the patients were elderly persons, there seems to be a need for more of geriatric work-out, comprehensive assessments, treatments and rehabilitation. This could be provided within these departments or in post-acute settings, specializing in geriatric medicine, until now non-existing in Estonia.

There is another potential for reducing costs while maintaining quality in health care – discharge planning with a nurse discharge planner or interdisciplinary team, see Johnson and McCargar (1999), Brooks (2001), Palmer et al (2001) and Noetscher et Morreale (2001). However, this calls for an implementation and use of a more comprehensive assessment, such as the Resident Assessment Instrument, of which the case-mix system of RUGs, used in this study, is a part.

6. Key points

- In Estonia many patients, particularly the older persons are being kept in hospital too long than it is actually required as regards their clinical needs.
- Comprehensive assessment of patients' needs is instead required for planning the structure of post-acute care and a reasonable use of resources to better develop care for older persons in Estonia.

REFERENCES

BROOKS, N.1. Yip PS, Law CK (2002). Assessment of the future resources and needs for hospitalization in Hong-Kong SAR. *Int J Health Plan Manage* 2002; 17:113—22.

CLARKE A, ROSEN R. (2001) Length of stay in community hospitals. *Nurs Stand*, 15, pp 33—8.

CLARKE, A. and ROSEN, R. (2001) Length of stay. How short should hospital care be? *Eur J Public Health* 11, pp 166—70.

COUGHLAN, T. and O'Neill, D. (2001) General hospital resources consumed by an elderly population awaiting long-term care. *Ir Med J*, 94, pp 206—8.

ESTONIAN nursing care master plan 2015 *Eesti hooldusravivõrgu arengukava* (2001)
http://www.sm.ee/est/HtmlPages/hooldusraviarengukava/$file/hooldusraviarengukava.pdf

EUROPEAN observatory on health care systems. Health care systems in transition (2000) Norway.

HANSEN, F. R., SPEDTSPERG, K. and SCHROLL, M. (1992) Geriatric follow-up of home visits after discharge from hospital: a randomized controlled trial. *Age Ageing*, 21, pp 445—50.

JOHNSON, J. E and MC CARGAR, P. A. (1999) Bilevel multidisciplinary discharge planning and length of stay. *Nursing Connections*, 12, pp 43—8.

MADSEN, F. F, NORSKOV, B and FROLUND, L. (2002) Length of stay in medical departments. Similarities and differences in Denmark. *Ugeskr Laeger*, 164, pp 2498—501 [in Danish].

MELIN, A. L. and BYGREN, L. O. (1992) Efficacy of the rehabilitation of the elderly primary health care patients after short-stay hospital treatment. *Med Care*; 30, pp 1004—15.

NIKOLAUS, T., SPECHT-LEIBLE, N., BACH, M., OSTER, P. and SCHLIERF, G. A. (1999) Randomized trial of comprehensive geriatric intervention in the care of hospitalized patients. *Age Ageing*, 28, pp 543—50.

NOETSCHER, C. M. and MORREALE, G. F. (2001) Length of stay reduction: two innovative hospital approaches. *J Nurs Care Qual*, 16, pp 1—14.

ONEN, F, ABIDI, H, SAVOYE, L, ELCHARDUS, J. M., LEGRAIN, S. and COURPRON, P. H. (2001) Emergency hospitalization in the elderly in a French university hospital: medical and social conditions and crisis factors

precipitating admissions and outcome at discharge. Aging (Milano);13, pp 421—9.

PALMER H. C., ARMSTEAD N. S., ELNICKI, D. M., HALPERIN, A. K., OGERSHOK, P. R., MANIVANNAN, S, HOBBS, G. R and EVANS, K. (2001) The effect of a hospital service with nurse discharge planner on patient care in an academic teaching hospital. *Am J Med* 111, pp 627—32.

PANIS, L. J., VERHEGGEN, F. W. and POP, P. (2002) To stay or not to stay. The assessment of appropriate hospital stay: a Dutch report. *Int J Q Health Care*, 14, pp 55—67.

SCHEIDER, D. P., FRIES, B. E, FOLEY, W. J., DESMOND, M and GORMLEY, W. J. (1988) Case mix for nursing home payment: resource utilization groups, version II. *Health Care Financ Rev*, Dec Spec No, pp 39—52.

VAUGHAN, B. and WITHERS, G. (2002) Bed-blocking. Acute distress. *Health Serv J*, 112, pp 24—7.

YIP, P.S. and LAW, C. K. (2002). Assessment of the future resources and needs for hospitalization in Hong-Kong SAR. *Int J Health Plan Manage*, 17, pp113—22.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1175—1188

THE FACTOR ANALYSIS IN RESEARCH OF THE SPATIAL DIFFERENTIATION OF POLAND'S ECONOMIC STRUCTURE

Anna Malina[1]

ABSTRACT

The subject of the study is the analysis of a spatial structure of the economy in the voivodships cross-section. In the studies we can widely use the methods of multidimensional statistical analysis. The paper presents the results empirical studies in which the method of factor analysis was applied. The analysis has allowed us to:

a) separate the main factors on the basis of an initially given set of primary characteristics. The factors can be interpreted as determinants of development,

b) rang voivodships to the identified main factors,

c) determine groups of voivodships similar in the development level as well as in the economic structure (identification of types of structure in the economy).

Key words: factor analysis, determinants of development, classification of multidimensional objects, identification of the economic structure types.

1. Introduction

Multidimensional statistics methods are widely applied in the comparative analysis of spatial entities (countries, regions, counties [*poviats*], and municipalities [*gminas*]). The methods enable examination of the relations between the entities (objects) that are characterized by means of specific sets of diagnostic features. Additionally, they make it possible to specify quantitative disproportions in respect of the level of development of the analysed entities, and to classify the entities into relatively uniform groups. In practice, the most

[1] Cracow University of Economics, Department of Statistics, 27 Rakowicka Street, 31-510 Kraków, Poland.

frequently used methods include concentration analysis methods, objects' linear arrangement methods, cluster analysis, and econometric modelling methods[1].

This paper contains the results of research of the spatial analysis of Poland's economic structure, the results being obtained on the basis of factor analysis.

Factor analysis consist of a number of methods and numerical procedures that allow the researcher to convert a given, primary set of correlated variables into a new structure of independent factors or principal components. The factors keep relatively large amount of the information contained in the primary variables, expressing, at the same time, new substantial content.

Some basic aspects of the application of factor analysis may be pointed out as follows:

a) it is a method of reduction of a set of primary features, replacing them with new uncorrelated variable (factors),
b) it is a method of determination of development determinants,
c) it is a method of arrangement and classification of multidimensional objects,
d) it is a method of graphic presentation of entities (objects) that are characterized within multidimensional spaces of features.

2. Theoretical assumptions to factor analysis

The key aim of factor analysis is to express a set of primary (standardised) variables in the form of the linear combinations of k common factors and m specific factors with the coefficients of, correspondingly, w_{jl} and w_j. The relation may be expressed as the following system of m equations[1]:

$$Z_j = \sum_{l=1}^{k} w_{jl} F_l + w_j u_j \quad (j = 1,...,m) \tag{1}$$

or as a matrix:

$$Z = W F + U. \tag{2}$$

where:

$Z = [Z_j]^T$ - transposed vector of standardised primary variables,

$F = [F_l^T]$ - vector of common factors $F_l (l = 1,...,k)$,

$W = [w_{jl}]$ - matrix of loadings of the l-th common factor of the j-th primary variable,

$U = [w_j U_j]^T$ - transposed vector of specific factors;

[1] See: for example: Hair and others (1992), Strahl (ed.), (1998), Malina, Zeliaś (1999), Zeliaś (ed.), (2000).

w_j - loading of the *j*-th specific factor,

m, k – number of primary variables and common factors $(k \leq m)$.

Factor analysis consists in determination of the w_{jl} parameters, which are called factor loadings, in such a way that the influence of the specific factors w_j is reduced to the greatest extent possible, for the common factors to take over. It is obtained by replacement of the R-element correlation in the matrix main diagonal with the common variance reserves $h_j^2 \ (j = 1,...,m)$. The common variance reserves h_j^2 are most often determined as the maximum correlation coefficient in respect to a given variable and the other primary variables:

$$h_j^2 = \max_i \{r_{ij}\} \quad i \neq j \tag{3}$$

As a consequence of the procedures, modified matrix $\widetilde{R}$ is obtained, the matrix elements being equal to:

$$\widetilde{r}_{ij} = r_{ij} \quad for \ \ i \neq j \tag{4}$$

$$\widetilde{r}_{ij} = h_j^2 \quad for \ \ i = j$$

Matrix $\widetilde{R}$ constitutes the basis for determination of factor loadings w_{jl} for the consecutive common factors $F_l \quad (l = 1,...,k)$. The procedure of factor loadings determination comes down to consecutively repeated steps, each step resulting in determination of the loadings of one common factor $F_l \quad (l = 1,...,k)$.

On the basis of matrix $\widetilde{R}$, the loadings of the first common factor F_1 are calculated, the share of the factor in the variance of the whole set of the features being maximised. In other words, the following function:

$$\psi_1 = \sum_{j=1}^{m} w_{j1} \tag{5}$$

is maximised by means of the Lagrange multipliers.

The procedure is repeated regarding the consecutive factors until the factors (altogether) explain the variance of the variables to the sufficient, required extent (for example, 90% or more)[2]. In practice to determine the number of factors we can use other approaches, for example, the "scree test criterion" or we can choose factors, for which eigenvalues are more than 1 [Hair, 1992, p. 246].

---

[1] See: Hair and others (1992), Dillion, Goldstein (1984), Jolliffe (1986), Grabiński (1992), Aczel (2000).

[2] According to the procedure assumed algorithm, the successive factors' contribution to the explanation of the total variance becomes smaller and smaller.

The examination of the conditional extremum of the function (5) by means of the Lagrange multipliers consists in determination of all non-zero and decreasingly ordered eigenvalues $\lambda_l \quad (l = 1,...,k)$ of matrix $\widetilde{R}$ and the corresponding eigenvectors $q_{jl} \quad (j = 1,...,m; l = 1,...k)$.

Eigenvalues λ_l constitute the basis for assessment of the total variance of the reserves, the variance being depleted by the following principal factors. The reserves may be expressed as percentages in the following way:

$$h_j = \frac{\lambda_j}{\sum_{j=1}^{m} \lambda_j}.100\%. \tag{6}$$

The sum of the consecutive *k* reserves determines the percentage share of the total variance that has been depleted by the *k* number of the first principal factors:

$$H_k = \sum_{j=1}^{k} h_j \tag{7}$$

Factor loadings w_{jl} represent the level of absorption of a given primary variable with particular factors. The loadings are calculated according to the following formula[1]:

$$w_{jl} = \frac{q_{jl}\sqrt{\lambda_l}}{\left[\sum_{j=1}^{m} q_{jl}^2\right]^{1/2}} \quad \begin{pmatrix} j = 1,...,m \\ l = 1,...,k \end{pmatrix}. \tag{8}$$

The matrix of factor loadings *W* enables substantial interpretation of the recognised principal factors. Factor $F_l \quad (l = 1,...,k)$ is identified by the set of the primary variables that are characterised by especially high[2] values (in respect of their modules) of the factor loadings $w_{jl} \quad (j = 1,...,m)$.

However, it is not always successful to differentiate separate groups of variables that identify particular principal factors. It often happens that the same primary variables are highly correlated with various factors, which makes their clear-cut identification difficult. In such situations, the so-called factor rotation[3] is carried out. The factor rotation is carried out in order to select such a position of

---

[1] It may be pointed out that the loading w_{jl} equals the correlation coefficient regarding variable Z_j and factor F_l (see: Grabiński, 1992).

[2] In practical applications, the variables whose factor loadings (in respect of their modules) are higher than 0.7 are most often selected for the sake of the identification of factors.

[3] The procedures most often applied for the factor rotation are the algorithms of „Verimax" and Quartimax" (see: Hair and others (1992), p. 235).

the common factors [$F_l \quad (l = 1,...,k)$] space axis that the corresponding matrix of the loadings enables relatively easy and unambiguous interpretation of the factors.

The subsequent stage of factor analysis is determination of the realisation of the successive principal factors for the entities (objects) under research. Common factors' values matrix *F* (of *n x k* dimensions) equals:

$$F = W^T Z, \tag{9}$$

where: W – factor loadings matrix; Z – matrix of primary variables' standardised values;
n – number of objects; k – number of differentiated factors.

Matrix *F* relating to the realization of the differentiated principal factors constitutes the basis for further statistical analysis, including econometric modelling, arrangement and grouping of the objects by the selected principal factors.

3. Results of empirical research

The basis for the performed research was the statistical data[1] that refer to Poland's economic structure by the voivodships (*provinces*) in 2000 year. The specification of initial variables are given in table 1.

In the set of the primary variables, four groups of describing variables have been differentiated:
a) municipalities' incomes and social infrastructure (symbol "*S*"),
b) ownership structure of economic entities (symbol "*P*"),
c) agricultural structure (symbol "*R*"),
d) level of the industry development (symbol "*Pp*").

The differentiated groups of variables refer to four basic sectors of the Polish economy in the spatial / territorial arrangement (by the voivodships). The selection of the variables was determined mainly by the subject-related reasons. Additionally, the variables to be examined were characterized by sufficiently high spatial variance (V>0.10).

[1] The input data to be the basis of the study are taken from the publication by the Central Statistical Office – *Statistical Yearbook of the Republic of Poland* and *Statistical Yearbooks of the Polish Voivodships*, GUS, Warszawa 2001.

Table 1. Specification of the primary variables of the social and economic structure of the voivodships

Item	Variable symbol	Name of the variable
1	S_1	Budgetary incomes of municipalities per capita (in PLN)
2	S_2	Municipalities' own incomes (% - share in the total incomes)
3	S_3	Number of housing spaces put into use 1000 residents
4	S_4	Hard-surfaced roads on the area of 100 km^2 (in km)
5	S_5	Subscribers of cable telecommunications lines per 1000 inhabitants
6	S_6	Net enrolment rate as for general secondary (high school) education (in %)
7	S_7	Population of cities and towns (%)
8	S_8	Number of physicians and dentists per 10 000 inhabitants
9	S_9	Unemployment rate (%)
10	P_1	Commercial companies (share in the number of all economic entities)
11	P_2	Foreign-interest companies (share in the number of all economic entities)
12	P_3	State-owned enterprises (share in the number of all economic entities)
13	P_4	Co-operatives (share in the number of all economic entities)
14	R_1	Arable land (as % of the areas in total)
15	R_2	Harvested wheat crops on 1 ha of arable land (in dt)
16	R_3	Harvested rye crops on 1 ha of arable land (in dt)
17	R_4	Harvested rape crops on 1 ha of arable land (in dt)
18	R_5	Harvested potato on 1 ha of arable land (in dt)
19	R_6	Harvested sugar beet on 1 ha of arable land (in dt)
20	R_7	Cattle per 100 ha
21	R_8	Pigs per 100 ha
22	Pp	Industrial production sold as per 1 inhabitant, in PLN (current prices)

Source: own study.

Primary variables matrix X, which includes the variables describing the above-mentioned sectors, may be presented in the following matrix block form:

$$X = \left[S \mid P \mid R \mid Pp\right]. \tag{10}$$

The initial data matrix X of the size 16 x 22 was standardized which resulted in matrix Z. On the basis of matrix Z the correlation matrix R was established as well as the factor loadings matrix W. (Because of the limited space of this paper the matrices X, Z and R has not been included in the present text.) Then the eigenvalues of the correlation matrix R were determined and the values of factor loadings for the following principal factors[1].

Table 2 contains eigenvalues λ_j of the correlation matrix in decreasing order, the variance reserves spent by the following principal factors $\hat{h}_j$ and cumulated reserves of total variation H_k, calculated according to formulas (6) and (7). Initially 11 principal factors were included which covered almost 90% of the total variation (variance). For the further analysis 6 first factors were assumed which

[1] All the numerical calculations were done with the pertinent statistic software "STATISTICA".

covered over 82% of information included in the primary values. The remaining factors are of lesser importance because eigenvalues λ_j pertaining to them are much smaller than 1 and none of them explains even 3% of the general variance.

Table 2. Values of parameters $\lambda_j, \hat{h}_j$ *i* H_k for the subsequent principal factors

Factor	λ_j	$\hat{h}_j$	H_k
1	5,6421	26,8675	26,8675
2	4,4869	21,3664	48,2340
3	2,6393	12,5683	60,8023
4	1,7903	8,5256	69,3279
5	1,3898	6,6182	75,9462
6	1,3493	6,4255	82,3717
---	---	---	---
7	0,5039	2,3998	84,7715
8	0,4049	1,9284	86,6999
9	0,2841	1,3530	88,0530
10	0,1336	0,6365	88,6896
11	0,0333	0,1587	88,8483

Source: own calculation.

In Table 3 there are the calculated values of factors loadings w_{jl} ($j = 1, ..., 22$; $l = 1,..., 6$) for the subsequent first principal factors. These loadings are coefficients of correlation between the initial variable and the principal factors and they constitute the basis for their content-related interpretation. The asterisk marks these high enough as far as the loadings value module is concerned. (the limiting values was assumed as $|r| \approx 0{,}70$). The assumption of the limiting value which is too low results in the over-identification of the factors because each of them has too many more and more different features. On the contrary the assumption of a too high limiting value may lead to relating the factors with single features which prevents their correct interpretation.

Table 3. Factor loadings matrix for the subsequent principal factors[1]

Variable	Factor F_1	Factor F_2	Factor F_3	Factor F_4	Factor F_5	Factor F_6
S1	0,8900*	-0,1040	0,2253	-0,0482	0,1212	0,1662
S2	0,9415*	-0,0252	-0,0345	0,0068	0,0819	0,1461
S3	0,2708	0,2307	0,1199	-0,0371	-0,1228	0,8970*
S4	-0,0375	-0,2895	0,6958*	-0,1497	-0,6108	0,0925
S5	0,8050*	0,3268	-0,0780	0,0838	-0,0616	0,3552
S6	0,1379	0,1845	-0,7922*	-0,3511	-0,0849	0,1513
S7	0,5471	0,1021	0,5059	-0,0823	0,3290	0,4732
S8	0,1921	0,2852	0,0627	-0,3618	-0,5864	0,5319
S9	-0,3102	0,0022	-0,1969	-0,1383	0,8052*	-0,0443
P1	0,9211*	0,1353	-0,0090	0,0762	-0,0335	-0,0457
P2	0,9382*	0,0528	0,0377	0,0398	0,1588	-0,0082
P3	-0,5807	0,0279	-0,2563	0,0772	0,1137	0,5291
P4	-0,2242	-0,1012	-0,8633*	0,2201	0,1775	-0,1694
R1	-0,0885	0,0469	0,4576	0,4630	-0,5658	0,3741
R2	-0,0812	-0,9408*	0,0085	-0,0664	0,1862	-0,1665
R3	0,3235	0,8462*	-0,0513	0,0475	0,1492	0,2947
R4	0,2260	-0,6996*	0,1321	0,0656	0,5485	0,1041
R5	-0,2324	0,2154	0,0354	-0,3921	-0,8152*	0,0342
R6	0,0818	-0,5263	-0,2683	0,6573	0,2843	0,1276
R7	-0,3775	0,3190	-0,0227	0,1263	-0,7544*	-0,0526
R8	0,1001	0,1745	0,0813	0,9449*	0,0152	-0,1052
P_p	0,6968*	0,1108	0,5384	0,2439	-0,2680	0,1181

Source: own calculation.

The results included in Table 3 show that the first principal factor (F_1), which covers nearly 27% of the total correlations, is fully identified by the variables describing the level and structure of municipalities' budgetary incomes (S_1, S_2), the presence of commercial companies and companies with foreign capital (P_1, P_2), the value of the industry output sold (P_p) and the number of telephone lines (S_5). This may be called an income-privatization factor.

The second principal factor (F_2), which covers over 21% of the total information is related to agriculture. It is identified by the variables R_2, R_3 and R_4 which characterize the effectiveness and structure of the plant economy (wheat, rye, rape).

[1] In the Table there are included factor loadings for isolated principal factors after doing the "Varimax" method rotation.

Factor F_3, which covers 12% of the total information is identified by the variables characterizing the local infrastructure (variable S_4) and the education of society (variable S_6).

Factors F_4 i F_5 should be also identified as the determinants of agriculture. They are identified by the variables which describe the animal economy (animal breeding) where factor F_4 is related to the breeding of swine and factor F_5 to breeding cattle. It must be mentioned here that factor F_5 is strongly related to the variable S_9, which describes the level of unemployment. This factor can be labeled as a determinant of structural unemployment in the regions dominated by agriculture.

The last distinguished factor F_6, which covers approximately 6.5% of information, is identified by one variable S_3, which describes the number of homes for the population of a given voivodship (region). Thus factors F_3 and F_6 may be described as the determinants of the socio-technical infrastructure.

The next step of the analyses requires determining the matrix F of the realization of isolated principal factors. Calculated according to (9) values of the next six principal factors for the units (voivodships) analyzed are included in Table 4.

The values of particular principal factors included in Table 4 may be the basis for the characteristics and description of the structure of the population of special units in relation to the isolated principal factors. In Table 5 there are voivodships arranged according to the decreasing values of the first three principal factors.

Table 4. Realization of principal factors for particular voivodships

Voivodship	Factor F_1	Factor F_2	Factor F_3	Factor F_4	Factor F_5	Factor F_6
Dolnośląskie	0,6529	-1,3039	-0,6936	0,3196	-0,8038	1,8358
Kujawsko-pom.	-0,3010	-0,5122	0,4464	-1,3422	1,3038	0,0643
Lubelskie	-1,1613	0,2041	1,2682	-0,4054	-1,5592	0,7899
Lubuskie	0,3838	-1,0981	-0,1197	1,2173	0,7742	-1,1290
Łódzkie	0,4973	1,1046	0,6903	0,4309	1,3816	1,1296
Małopolskie	-0,4546	1,1110	-0,8204	-0,3465	-1,2942	-0,6645
Mazowieckie	2,3836	0,9181	0,9706	-0,4347	-1,4945	-0,5083
Opolskie	-0,3763	-1,3489	-0,5330	-1,6681	-0,0916	1,1361
Podkarpackie	-1,5001	0,1723	-1,0298	0,3063	-0,7694	-1,3157
Podlaskie	-0,5724	1,3078	0,9462	0,7171	0,2892	-0,0686
Pomorskie	0,9184	-0,5514	0,0355	0,2818	-0,3696	-0,9659
Śląskie	0,9106	0,9135	-0,1509	0,4907	1,1526	0,8651
Świętokrzyskie	-1,2696	0,3198	-0,5460	0,3067	0,0562	0,7970
Warmin-mazur	-0,7438	-0,9619	0,3798	1,1456	0,6543	-0,7386
Wielkopolskie	0,2938	0,3601	0,3239	-2,1402	1,0664	-1,3580
Zach-pomorskie	0,5392	-1,0050	0,7406	1,1209	-0,2961	0,1309

Source: own calculation.

Table 5. Arrangement of voivodships according to the values of the first three principal factors

Factor F_1		Factor F_2		Factor F_3	
Voivodship	*Value of a factor*	*Voivodship*	*Value of a factor*	*Voivodship*	*Value of a factor*
Mazowieckie	2,3837	Podlaskie	1,3079	Lubelskie	1,2683
Pomorskie	0,9184	Małopolskie	1,1110	Mazowieckie	0,9706
Śląskie	0,9107	Łódzkie	1,1046	Podlaskie	0,9462
Dolnośląskie	0,6529	Mazowieckie	0,9182	Zach-pomorskie	0,7407
Zach-pomorskie	0,5392	Śląskie	0,9136	Łódzkie	0,6903
Łódzkie	0,4973	Wielkopolskie	0,3601	Kujawsko-pom.	0,4464
Wielkopolskie	0,3838	Świętokrzyskie	0,3198	Warmińsko-maz	0,3798
Lubuskie	0,2938	Lubelskie	0,2041	Wielkopolskie	0,3239
Kujawsko-pom	-0,3010	Podkarpackie	0,1723	Pomorskie	0,0355
Opolskie	-0,3763	Kujawsko-pom.	-0,5122	Lubuskie	-0,1198,
Małopolskie	-0,4546	Pomorskie	-0,5515	Śląskie	-0,1509
Podlaskie	-0,5724	Warmińsko-maz	-0,9619	Opolskie	-0,5331
Warmińsko-maz	-0,7438	Lubuskie	-1,0981	Świętokrzyskie	-0,5460
Lubelskie	-1,1614	Zach-pomorskie	-1,0050	Dolnośląskie	-0,6936
Świętokrzyskie	-1,2696	Dolnośląskie	-1,3039	Małopolskie	-0,8204
Podkarpackie	-1,5001	Opolskie	-1,3489	Podkarpackie	-1,0298

Source: own study.

One must notice that when factor F_1 is of high value factor F_2 is of low value and the other way round. (cf. Table 5). The voivodships which have high value of factor F_1 are characterized by high level of municipalities' budgetary incomes per head, higher level of privatization and more developed socio-technical infrastructure. These voivodships are: Mazowieckie, Pomorskie, Śląskie, Dolnośląskie, where Mazowieckie has got especially high value of this factor (2,3837). On the other hand the voivodships of the lowest values of factor F_1 are the voivodships of south-eastern and eastern Poland (Podkarpackie, Świętokrzyskie, Lubelskie, Podlaskie, Warmińsko-mazurskie). On the whole these are the regions of agricultural character, poor infrastructure and low level of economic development. These voivodships have quite high values of factor F_2 (with the exception of Warmińsko-Pomorskie Voivodship) — see Table 5.

While analyzing the arrangement of voivodships according to the second isolated factor F_2 , we can notice that the highest values of factor F_2 are in voivodships: Podlaskie, Łódzkie and Małopolskie (1,1—1,3), average values in voivodships: Lubelskie, Podkarpackie, Świętokrzyskie, and the lowest in the

voivodships of western Poland: Lubuskie, Zachodnio-pomorskie, Dolnośląskie oraz Opolskie.

Factor F_3 has the least uniform character, which does not allow an unequivocal, content-related interpretation. In fig. 1 there is a diagram of dispersion of voivodships in relation to the first two principal factors.

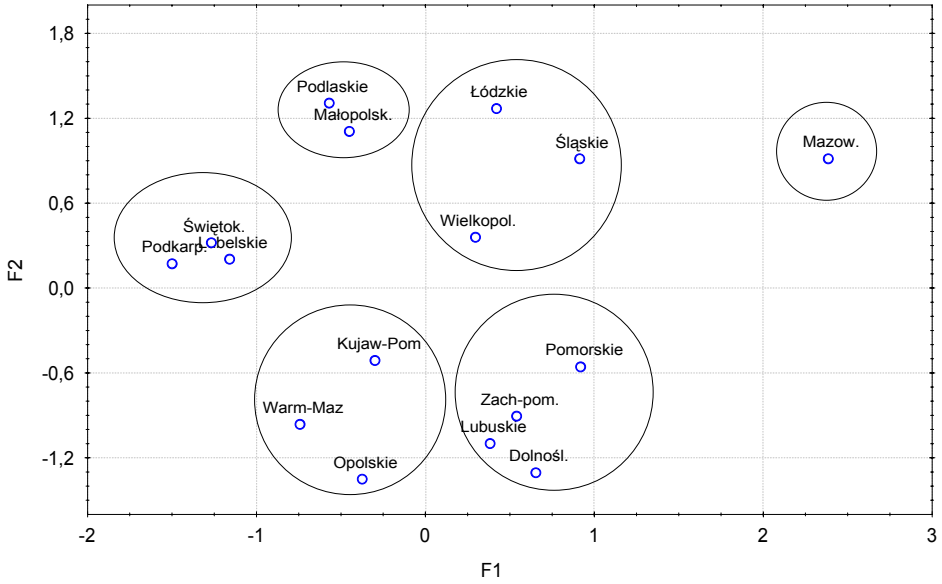


Fig. 1. Dispersion of voivodships in relation to the first two principal factors.

On the basis of the realization of the first two common factors (F_1 and F_2), which together cover almost 50% of the total number of variants considered in variables analysis, the voivodships have been divided into groups of similar economic structure. The particular groups are included in Table 6 (cf. Fig. 2).

Table 6. Classification of voivodships according to the value of the first two principal factors

Group	Voivodship	Average value of a factor		Region type (structure)
		F_1	F_2	
No. 1	Mazowieckie	2,384	0,918	Well developed (diversified)
No.2	Łódzkie, Śląskie, Wielkopolskie	0,667	0,793	Developed (diversified)
No.3	Dolnośląskie, Zach-pomorskie, Pomorskie, Lubuskie	0,623	–0,989	Average developed (diversified)
No.4	Małopolskie, Podlaskie	–0,513	1,209	Average developed (agriculture dominating)
No.5	Lubelskie, Podkarpackie, Świętokrzyskie	–1,310	0,266	Poorly developed (agricultural profile)
No.6	Opolskie, Kujawsko-pomorskie, Warmińsko-mazurskie	–0,474	–0,941	Poorly developed (agriculture dominating)

Source: own study.

The above analysis clearly shows that Mazowieckie Voivodship is different (in positive sense) from all the others according to factor F_1 connected with municipalities' budgetary incomes and privatization. The value of F_2 in the case of this voivodship is also high. Thus Mazowieckie Voivodship may be described as an example of a developed type of an economic region. The voivodships of the second group (Śląskie, Wielkopolskie, Łódzkie) may be considered as developed of diversified economic structure too. The values of the first two factors are positive and close to 1. The voivodships of the first and second groups are characterized by the *expansiveness* of their economic structure. Furthermore, we should notice that voivodships Pomorskie, Zachodnio-pomorskie and Dolnośląskie constitute a uniform group with relatively high value of factor F_1 and at the same time low value of factor F_2. This group may be described as an average developed region. The voivodships Małopolskie and Podlaskie have high values of F_2 and relatively low values of factor F_1 (cf. Fig. 1). The voivodships of the fifth group (Podkarpackie, Świętokrzyskie i Lubelskie) have very low values of factor F_1 (the average value of this factor is 1,31) and the values of factor F_2 close to zero. Thus this group is characterized by the low level of social development and poorly developed infrastructure. The similar situation pertains to voivodships of the sixth group (Opolskie, Kujawsko-pomorskie, Warmińsko-mazurskie), the values of both factors are below zero and the average value of factor F_1 is higher than in the fifth group (cf. Table 6). The voivodships of the last two groups (with the exception of Kujawsko-pomorskie) have been described as regions of *crisis-causing* economic structure.

4. Conclusions

The present paper includes the results of comparative analysis of voivodships obtained by the means of factor analysis.

Employment of this method has allowed isolating the basic principal factors from the given set of variables describing the socio-economic structures of voivodships. These factors may be interpreted as determinants of development of regions. The conclusion derived from the analysis is that in the comparative analysis of Polish regions the most important are two factors: F_1 related to municipalities' budgetary incomes and privatization and F_2 related to the structure of agriculture.

The values of isolated principal factors were the basis for arrangement and classification of voivodships into uniform typological groups. According to the values of the first two principal factors the voivodships were ascribed to groups characterised by the similarity of economic structure.

REFERENCES

ACZEL A.D. (2000), Statistics in Management (In Polish), Warszawa: PWN.

DILLION W. R., GOLDSTEIN M. (1984), Multivariate Analysis: Methods and Applications, New York: Wiley.

GRABIŃSKI T. (1992), Factor Analysis in Spatial research [in:] Spatial Research of Market and Consumption. Methodological Guide (editor: S. Mynarski), (In Polish), Warszawa: PWN.

HAIR J. F., ANDERSON R. E., TATHAM R. L., BLACK W. C. (1992), Multivariate Data Analysis with Readings, 3-rd edition, Macmillan Publ. Company, New York.

HARMAN H. H. (1970), Modern Factor Analysis, University of Chicago Press, Chicago.

JOLLIFFE I. T. (1986), Principal Component Analysis, Springer-Verlag, New York.

MALINA A. (2002), Application of Taxonomy in Multi-criterion Comparative Analysis of Socio-economic Structures, [in:] Time-spatial Modelling and Predicting of Economic Phenomena, (editor: A. Zeliaś), (In Polish), Cracow University of Economics Publishers, Kraków.

MALINA A., ZELIAŚ A. (1999), Taxonomic Analysis of the Spatial Differentiation of the Living Conditions in Poland,[in:] Economic and Social

Statistics in Transition.Proceedings of the Ukrainian, Polish and Slovak Scientific Seminar, Kiev .

STRAHL D. (editor), (1998), Taxonomy of Structures in Regional Research (In Polish), Science Papers Economic Academy in Wrocław, Wrocław.

ZELIAŚ A. (editor), (2000), Taxonomic Analysis of Spatial Diversification of the Standard of Life in Poland in Dynamic Approach (In Polish), Cracow University of Economics Publishers, Kraków.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1189—1195

REPORTS

CELEBRATING STATISTICS
INTERNATIONAL CONFERENCE IN HONOUR OF SIR DAVID COX ON THE OCCASION OF HIS 80th BIRTHDAY

Neuchâtel, 14—18 July 2004

Neuchâtel, the capital of Canton of Neuchâtel in Switzerland, a notable seat of culture and learning, with a university that was founded in 1838, was a place of an international conference in honour of Sir David Cox on the occasion of his 80th birthday. The conference was organized by a Statistics Group of the University of Neuchâtel with the collaboration of École Polytechnique Fédérale de Lausanne and Mainz University. Members of the scientific committee were: Sir David Cox, Anthony Davison, Yadolah Dodge and Nanny Wermouth.

Sir David Cox, a world-famous statistician, was educated at St John's College, Cambridge (MA), and the University of Leeds (PhD, 1949). After working in industry, he held academic posts at Cambridge and Birkbeck College, London. In 1966 he became Professor of Statistics, and in 1970 Head of the Mathematics Department, at the Imperial College of Science and Technology. He is a former Warden (1988—94) and an honorary Fellow of Nuffield College, Oxford. He is also a Fellow and former Member of Council of the Royal Society. He was President of the Bernoulli Society (1979—81), the Royal Statistical Society (1980—82) and the International Statistical Institute (1995-97). He holds a number of awards, including honorary Fellowships at St John's College, Cambridge, and the British Academy, the Guy Medal (in Silver and in Gold) from the Royal Statistical Society, and the degree of DSc from 18 universities, Neuchâtel being one of them. He has also been awarded the Weldon Memorial Prize, University of Oxford; the Kettering Prize; Gold Medal for Cancer Research and the Max Planck Forschungspreise. He is a Foreign Honorary Member of the US National Academy of Sciences. He was also editor of Biometrika from 1966 to 1991. He has authored or co-authored sixteen books and roughly 250 papers from statistical theory, methods and applications. His work on the proportional hazards regression model is one of the most-cited and most influential papers in modern statistics. Sir David Cox's current affiliation is Department of Statistics, University of Oxford.

The scientific part of the conference was preceded by an opening ceremony in which leading Neuchâtel society members took part: M. Dubois — Dean of the Faculty of Economics, H. Nägeli — Co-Rector of the University of Neuchâtel, J-J. Clémençon — Chief of the University Office of the Cantonal Gouvernment and B. Soguel — Chief of the Economics Office of the Cantonal Gouvernment.

The conference in honour of Sir David Cox gathered over 90 participants from all over the world: from Argentina (3 participants), Austria (1), Belgium (1), Brazil (1), Canada (5), China (3), Cyprus (1), Czech Republic (1), Denmark (2), France (2), Germany (2), India (1), Iran (1), Israel (3), Italy (9), Japan (1), Mexico (1), Poland (1), Slovak Republic (1), South Africa (4), Spain (2), Sweden (2), Switzerland (11), Netherlands (1), Turkey (3), United Kingdom (20), USA (9). The conference scientific programme included 11 Invited Talks presented during 45 minutes long lectures. The subjects of the talks comprised wide area of modern statistics, from theory, to computations and applications in medicine, biology, environmental studies, social studies and finance. There were also 37 contributed papers and 8 posters introduced different projects and research areas within theory and applications of statistics. A short discussion took place right after each lecture, but they often continued in the break and also lunchtime.

SCIENTIFIC PROGRAMME OF THE CONFERENCE

Invited Speakers:

- Ole Eiler Barndorff-Nielsen (University of Aarhus, Denmark) in his talk *On Some Concepts of Infinite Divisibility and their Roles in Turbulence, Finance and Quantum Stochastics* reviewed a few aspects and ramifications of Lévy Theory (the theory of infinite divisibility, Lévy processess, Lévy bases etc.) and its applications.
- Sarah C. Darby (CCTSU Oxford, UK) gave a talk *The Early Breast Cancer Trialist's Collaborative Group: a Brief History of Results to Date* in which some of the results from the most recent cycle of the Early Breast Cancer Trialist's Collaborative Group (EBCTCG) were summarized, concentrating on the results from approximately 38,000 women who have been randomised in trials of "local" therapies, that is trials of radiotherapy versus no radiotherapy, more surgery versus less surgery, and more extensive surgery versus radiotherapy plus less extensive surgery. Breast cancer is the commonest cancer among women, and each year is responsible for an enormous number of deaths. Mortality rates from breast cancer have now started to decrease in many parts of the world due, at least in part, to the work of the EBCTCG which carries out collaborative meta-analyses on the treatment of any aspect of early (operable) breast cancer.

- David Firth (University of Warwick, UK) in his talk *Some Topics in Social Statistics* took a rapid tour through some areas of statistical research which are of particular relevance to at least some types of social science. Remarks were made on: multi-level (or hierarchical or mixed-effects) models; the use of statistical models for prediction (e.g., small-area estimation based on large-scale surveys); the use of statistical models as measuring instruments, for example in the study of social fluidity, health inequality or local deprivation; the role of sampling weights; and various other topics.
- Peter Hall (Australian National University, Canberra, Australia) presented a talk *On Nonparametric Statistical Methods* in which the contemporary context of nonparametric statistics was discussed. To some statisticians, the evolution of nonparametric methods over the last quarter-century has exemplified a movement of statistical science away from its mathematical roots, towards a far more experimental discipline closely allied with computing. One can also take the contrary view, however: that the demands of mathematically elucidating the complexities of nonparametric methods have moved the field particularly close to mathematics.
- Valerie Isham (University College London, UK) spoke about *Stochastic Models for Epidemics*. Human infections such as influenza, malaria and HIV are still major world-wide causes of morbidity and mortality. Understanding the dynamics of the spread of an infection disease brings possibilities for its control and prevents a large-scale epidemic. A clear understanding of the basic theory of epidemic models is a prerequisite for this and, equally, for informing the development of the successful vaccination policies against a wide range of infections that are routinely used around the world. In her talk, Valerie Isham focused on stochastic epidemic model. She briefly reviewed their historical background and described some of the topics that preoccupied researchers over the last fifteen years or so, looking at progress that has been made and identifying continuing challenges. Concentration was made on model structure and general modelling issues of infections for animal/human hosts.
- Peter McCullagh (University of Chicago, USA) spoke about *Exchangeability and Regression Models*. The definition of an exchangeable regression process asserts only that two finite samples of units having the same x-values also have the same response distribution. In this form regression exchangeability implies lack of inference from other units and, in a certain mathematical sense, absence of unmeasured confounders. It does not imply independence of the response on distinct units. All classical linear and generalized linear models are regression exchangeable, as are some non-classical models. The consequences of regression exchangeability were discussed. Extensions to more complicated situations in which a block factor or other relationship is defined on the units were also considered.
- Nancy Reid (University of Toronto, Canada) in her talk *Theoretical Statistics and Asymptotics* begun with Cox and Hinkley's *Theoretical Statistics* (1974),

which was a uniquely modern treatment of the foundations and theory of statistical inference. Since the time of its publication there has been considerable development of the theory of inference, especially in the area of higher order likelihood asymptotics. There have also been very dramatic advances in the implementation of Bayesian methodology, particularly with the development of sampling methods for computing posterior distributions. Nancy Reid considerd how these developments might inform the current study of theoretical statistics, with particular reference to selected examples from *Theoretical Statistics.*

- Brian D. Ripley (University of Oxford, UK) presented a talk *How Computing Has Changed Statistics*. The pace of change in computing resources appears relentless. The challenge is to find innovative ways to use computation to improve our statistical inference, for example by using more realistic models and improving approximations, without data dredging, over-analysing the data to hand. The problems how computational complexity interacts with the growth in computing resources and how software is crucial to the use of new statistical methods were widely discussed by the author.
- Ignacio Rodriguez-Iturbe (Princeton University, USA) in his talk *Stochastic Soil Moisture Dynamics and Vegetation Response* briefly summarized the results of a stochastic soil moisture model that describes the dynamics of the soil water balance via a nonlinear differential equation driven by a state-dependent marked Poisson process. The analytical results of the steady-state probability density function and the crossing analysis provide a synthetic representation of plant response to hydro-climatic forcing that well describes vegetation in water-controlled ecosystems. [The paper together with Amilcare Porporato as co-author].
- Neil Shephard (University of Oxford, UK) gave a talk *Econometrics, Statistics and Financial Volatility*. In a very interesting talk the author compared a parametric and semiparametric analysis of volatility estimation and jump detection in a problem drawn from financial econometrics.
- Scott L. Zeger (The Johns Hopkins University, USA) in his talk *A Cox Model for Biostatistics of the Future* briefly reviewed exciting biomedical and public health challenges capable of driving statistical developments in the next decade. He discussed statistical models and model-based inferences central to the CM approach contrasting them with computationally-intensive strategies for prediction and inference advocated by Breiman and to more traditional design-based methods of inference. He discussed the hierarchical model as an example of the future challenges and opportunities for model-based inference. Recent examples from genetics were used to illustrate the ideas. Two other topics: causal inference and statistical computing were also discussed. [The paper together with Peter J. Diggle and Kung-Yee Liang as co-authors].

The work of Sir David Cox was often directly relevant to the specific topics discussed by invited speakers.

Plenary contributed paper session (Chair: Nancy Reid)

- P.J. Solomon: Planning Microarray experiments
- D. Oakes: On Collapsibility of Time-Scales in Survival analysis
- R.C.L. Wolff: Chaos and Asymptotics: Towards Exact Inference for Dynamical Systems
- P. Sprent: The Cauchy Distribution: an Enigma with Several Variations
- A. Azzalini, T. Rosolin and N. Torelli: Clustering via Nonparametric Density Estimation

Contributed paper session 1 (Chair: David Firth)

- J. Sergeant: Social Differentials in Health: The Relative Index of Inequality
- A. Gottard, C. Rampichini: Chain Graphs for Multilevel Models
- G.M. Marchetti and E. Stanghellini: On Testing the Existence of One Unobserved Confounder
- B. Kowalczyk: Sampling on Successive Occasions: Estimation of a Population Ratio

Contributed paper session 2 (Chair: David Hinkley)

- G.A. Young: Bootstrap and the Fisherian Proposition
- P. Northrop: Estimating the Parameters of Rainfall Models Using Maximum Marginal Likelihood
- D.A.S. Fraser and N. Reid: Assessing a Vector Parameter
- M. Akahira: The Concept of Large-Deviation Efficiency and Its Applications

Contributed paper session 3 (Chair: Peter Hall)

- R.J.Samworth: Bagging Nearest-Neighbour Classifiers
- C.C. Taylor: Bagging Kernel Learners
- S. Oosthuizen and S.J. Steel: Stepwise Input Variable Selection for Support Vector Classifiers
- G. Schoier: On Spatial Data Mining and Clustering methods: a Proposal
- G. Brunet: Hierarchical Divisive Method with Parallel Programming

Contributed paper session 4 (Chair: Patty Solomon)

- H. Khamis: Bias of the Cox Model
- C. Cigsar and A.D. Akkaya: Goodness of Fit Statistics Based on Censored Samples: Extreme Value Distribution
- Q.-G. Wu and J.-H. Lu: Sampling Inspection of Reliability in (Log)Normal Case with Type I censoring
- Y. Tian and G.-Y. Li: Saddle-Point Approach to Inference in Sensitivity Experiments Under Logistic Response Distributions

Contributed paper session 5 (Chair: Rodney Woolf)

- G. Kauermann: Some Aspects of Penalized Spline Smoothing and Generalized Linear Mixed Models
- A.W. Roddam: Meta-Analysis of Diagnostic Tests with an Application in Prostate Cancer Screening
- E. A. Gehan: Some Applications of Bayesian Methods in Cancer Clinical Trials

Contributed paper session 6 (Chair: Anthony Davison)

- F. Azizi: Household Division of Labour: Developing the Rasch Model
- D. Larocque and M. Labarre: A conditionally Distribution-Free Multivariate Sign Test for One-Sided Alternatives
- W.J. Reed: Estimating Historical Forest Fire Frequency from Multiple Scar data
- A. Giovagnoli and H.P. Wynn: Stochastic Orderings and Measures of Agreement for Discrete Random variables

Contributed paper session 7 (Chair: Adelchi Azzalini)

- Y.E. Yilmaz and A.D. Akkaya: Analysis of Variance and Linear Contrasts in Experimental Design with Symmetric Non-Normal Error Distributions
- D. Uys and S. Steel: Influential Cases when the C_p Criterion is Used for Variable Selection in Multiple Linear Regression
- E. Moreno: Calibrating *p*-Values in Linear Models
- M.I. Araujo, R. Cleroux, M. Fernandes, A. Lazraq and B. De B. Pereira: Alternative Procedures to Discriminate Non-Nested Multivariate Linear Regression Models

Contributed paper session 8 (Chair: Valerie Isham)

- S. Fruehwirth-Schnatter and H. Wagner: Gibbs Sampling for State Space Modelling of Time Series of Counts
- K. Fokianos: Some Further Results on Regression Theory for Time Series of Counts
- G. Weiss: Multi-Class Queuing Networks with Infinite Virtual Buffers
- M. Molina, M. Mota and A. Ramos: Bayesian Inference for Bisexual Processes with Population-Size Dependent Mating

Poster session

- E. Altpeter and B. Vidondo: Reproducibility of Medical Diagnostic Tests: An Application
- J.A. Visek: Least Weighted Squares for Dynamic Specification
- I. Zezula: Some Remarks on the Extended Growth Curve Model

- K. Bingavanga: A Simple Method of Calculating Cumulants of Polynomials
- E. Diaz-Frances: The Estimation of Animal Abundance Using the Binomial Models
- V. Bagdonavicius, F. Haghighi and M. Nikulin: The Application of Degradation Model in Survival Analysis and Reliability
- M. Giacalone: Effects of Multicollinearity on the Linear Regression Model Estimations
- K.O'Rourke: Especially Why Random Effects and Whatever for?

Organizers of the conference also assured eventful social programme, including aperitif at the Neuchâtel Castle, offered by the Gouvernment of the Canton of Neuchâtel, Gala Dinner at the Hotel du Peyrou in honour of Sir David Cox, a boat trip and a visit at Latenium, a reception from the authorities of Neuchâtel at the museum of Arts and History. Excellent scientific programme, stimulating session and after session talks, perfect organization, eventful social programme and most of all friendly atmosphere — made the conference days most enjoyable and valuable.

Prepared by Barbara Kowalczyk
Warsaw School of Economics
bkowal@sgh.waw.pl

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1197—1203

27th CIRET Conference: Review of Selected Papers,

Warsaw, Poland, 15—17 September 2004

The 27th CIRET Conference was held in Warsaw from 15th to 17th September 2004. Organisation of the Conference was charged to the *Research Institute of Economic Development of the Warsaw School of Economic*s in cooperation with the *Central Statistical Office of Poland and the National Bank of Poland*.[1] Such meetings are organized in one of the member countries every other year. Next meeting is planned for autumn 2006 in Rome.

CIRET is an abbreviation from the ***Centre for International Research on Economic Tendency Surveys***. It was created in 1953. The present president of CIRET is director of the *Swiss Institute for Business Cycle Research* (*Konjunkturforschungsstelle der ETH Zűrich* — KOF) — Professor Bernd Schips, vice-president — Gernot Nerb from the *Munich Institute for Economic Research (Institut fűr Wirtschaftsforschung* — Ifo). Organisation gathers 53 individual members and 35 corporate members.

Business tendency surveys it is a special system of surveys on enterprises and consumers aiming at collection of opinions of particular groups of respondents on present and future situation in the nearest months. These are qualitative surveys e.g. providing general information on tendencies of development in the economy and feelings of consumers so they do not aim at gathering of the definite quantitative information (of which the source is the official short term statistics). This type of surveys is widely spread in Europe and outside. For many years European Union and OECD has been introducing *Joint Harmonised Program of the Business* and *Consumer Tendency Surveys* in the member states with the use of the harmonised set of questions for particular areas of the economy and consumers, elaborating common definitions of indicators, publishing results for particular countries or groups of countries.

Planned meetings took place in parallel sessions. Their main subjects were:

- Business tendency surveys, including investment survey: analysis of surveys results, new methods and other chosen subjects;
- Consumer surveys: analysis of results, new methods;
- Business tendency indicators — analysis and new indicators;
- Special surveys.

[1] The Polish version of this report was published in "Wiadomości Statystyczne" No 11/2004

Because there were almost 90 presentations prepared for the conference only chosen papers from particular sessions are presented below.

Business tendency surveys

Business tendency surveys were the dominant subject of the conference. Papers prepared by participants were presented in eleven sessions during which the results of different kinds of business tendency surveys (already implemented or new ones) were presented as well as their use in micro- and macroeconomic analysis of present economic situation and its short term forecasts (theoretical and practical ones). Such surveys comprise manufacturing industry, construction, retail trade and service enterprises (including financial sector) as well as surveys on investment activity. In many countries they are conducted according to the Joint Harmonised Programme of Business and Consumer Surveys passed by the European Union and OECD what means that obtained results are fully comparable on international scale.

First group comprised: six sessions dedicated to the presentation of different analytical works using results of the business tendency surveys, four sessions intended for presentation of new methods and new surveys, and one session in the course of which authors presented other subjects connected with the business tendency surveys.

Results of the business tendency surveys conducted by the Polish institutes were presented by representatives of the *Warsaw School of Economics* (WSE), the *Central Statistical Office* and the *Poznań Academy of Economics*. On the first day of the conference Emilia Tomczyk from *Research Institute of Economic Development* of the WSE reflected on what information is contained in replies formulated by the Polish businessmen conducting manufacturing activity (*Are Expectations of Polish Industrial Enterprises Rational?*). Monika Dedys and Barbara Tarnowska from the same Institute presented paper on possibilities of use of binary hidden Markov models in analysis of results of business tendency surveys (*Binary Hidden Markov Models in Analysis of the Results of Business Surveys)*. Authors of the third paper from the *Research Institute of Economic Development* — Elżbieta Adamowicz, Sławomir Dudek and Konrad Walczyk — were considering possibilities of use of qualitative business tendency surveys results in short term forecasting (*The Usefulness of Business Surveys Data for Short-term Forecasting. Raw Data vs. Seasonally Adjusted and Smoothed One*). A group of authors from the *Central Statistical Office of Poland* — Hanna Sękowska, Magdalena Święcka, Katarzyna Walkowska and Izabella Zagoździńska presented results obtained from the business tendency surveys of investments in manufacturing industry and construction *(The Assessment of the Investment Activity of Enterprises based on the Business Tendency Surveys*). During the separate session on chosen problems of business tendency surveys the group of researchers from Poznań presented their achievements in surveys on

financial sector institutions. Results of the surveys on banks were presented by Józef Garczarczyk and Robert Skikiewicz (*Forecasting Business Surveys Data in Diagnosis and Forecasting Situation of Polish Economy*). Similar considerations on insurance institutions were presented by Józef Garczarczyk, Marek Mocek and Iwona Olejnik (*Business Tendency Surveys in Diagnosing and Forecasting the Market of Insurance Services in Poland*).

France is one of the countries where — apart from Italy and Germany — qualitative business tendency surveys have the longest tradition in Europe. On the conference first of the papers prepared in the *French National Institute of Statistics and Economic Research* (*Institut National de la Statistique et des Etudes Economiques* — INSEE) concerned the assessment of the quality of information on competition based on the results of the quarterly business tendency surveys (Patrick Aubert and Marie Leclair — *How informative are Firms' Statements about Competitiveness?*). Second French paper written by Antoine Naboulet and Sebastien Raspiller referred to investment survey and its authors considered what factors influence the decisions of entrepreneurs on undertaking investment activity (*Determinants of Investment Decisions: What can we learn from Business Investment Surveys?*). François Bouton and Hélène Erkel-Rousse showed the possibilities of use of business tendency surveys results of particular areas of economy in French short term macroeconomic forecasts (*Sectoral Business Surveys as an Aid to Short-Term Macroeconomic Forecasting*).

Italian researchers from the *Institute of Researches and Economic Analysis (Istituto di Studi e Analisi Economica* — ISAE) prepared for the conference three papers referring to business tendency surveys. They concerned implementation of new methods in conducted surveys: in sampling method and in the system of processing and grossing up of data. Works conducted with reference to manufacturing industry surveys were presented by Marco Malgarini, Patrizia Margani and Bianca Maria Martelli (*Re-engereneering the ISAE Manufacturing Survey*), to services survey — by Bianca Martelli and Gaia Rochetti (*The ISAE Market Services Survey*), and to investment — by Tatiana Cesaroni, Marco Malgarini and Gaia Rocchetti (*Quality Control and Aggregation Methods in the ISAE Investment Survey*).

Except the countries mentioned above, the significant output in the field of the business tendency surveys in Europe have also scientist from Finland. In two Finnish papers results of the prognostic works with use of the business tendency surveys were presented. The first one presented by Jouko Kangasniemi, Kari Takala and Pekka Tsupari concerned method of the forecasting of the manufacturing production (*Forecasting Finnish Manufacturing Production. Evaluation of the Predictive Power of Business Tendency Survey Results*). Presentation of Samu Kurri and Kari Takala showed the possibilities of use of barometers models in short term GDP forecasting (*Forecasting GDP with Barometers Models in Finland*).

Central and Eastern European countries were represented — apart from hosts of the Conference — also by representatives of Hungary, Czech Republic, Russia and Ukraine. In Hungary business tendency surveys are conducted by private the *Institute for Economic Research (GKI Gazdaságkutató Rt.)*. Scientists from this institute — Gabor Papanek, Raymund Petz and Zoltan Sulok — presented paper on forecasts of the Hungarian economy (*Reliability of economic forecasts in Hungary*), and Judit Adler, Judit Barta, Tamas Kopik and Raymund Petz analysed results of the business tendency survey in service sector (*The Contradiction between Survey Results in the Service Sector and Service Sector Situation. Causes and solutions*).

Czech Republic was represented by the expert from the *National Bank of Czech Republic* Jakub Fischer who showed in his paper how results of the business tendency surveys could be used in prognostic works concerning Czech economy (*Possibilities of Predicting with Using Business Tendency Survey*).

In his paper the representative of the *Russian Institute for the Economy in Transition* Siergiej Tsukhlo compared information obtained from the surveys conducted by the independent private scientific institutes with information the source of which is the official quantitative statistics (*Business Surveys and Official Statistics in Russia. Which one is better?*). Paper from Ukraine was presented by Maryna V. Pugachova and Wiktor A. Golovkov. The subject of their presentation was the use of the business tendency surveys results in macroeconomic analysis (*Using the Business Tendency Survey Data for Macro Economic Analysis*).

Consumer surveys

Consumer surveys are conducted in particular countries by different institutes, very often by private institutes or centres for public opinion poll. In this field there were much less papers than in the field of business tendency surveys and business tendency indicators. Results of the consumer surveys were presented in two thematic groups: first two sessions concerned analysis of the surveys results, the third one — new methods used in this type of surveys.

Polish experiences in this field were presented by Barbara Kowalczyk from the *Warsaw School of Economics* in the paper on the use of the qualitative data in sampling (*Some Application of Qualitative Data in Survey Sampling*).

Italian paper from ISAE written by Solange Leproux, Marco Malgarini and Patrizia Morgani concerned the ability of consumers to assess their present and future situation in context of their real expenditures (*Consumer Sentiment and Households Expenditures in Italy: A Disaggregated Analysis*). Second paper from Italy prepared by Maurizio Bovi also concerned the quality of answers of consumers (*Consumer Surveys and Reality*).

Representatives of the French institute INSEE Hélène Poncet considered how the harmonisation of the French surveys with the European Programme

influenced their quality (*The Consequences of European Harmonization on the French Consumer Survey*).

Representatives of Finland Pertti Kangassalo and Kari Takala presented the use of the results of the consumer tendency surveys in forecasting works on size of expected inflation (*Measuring the Usefulness of Consumers' Inflation Expectations in Finland*).

Business tendency indicators

This set of subjects comprised papers of which authors presented their theoretical and practical works concerning composite and leading indicators.

Use of composite indicators (based on simple indicators) is aimed at description of economic situation of the whole economy of the country or sectors of economy (for instance manufacturing industry, construction etc) in a synthetic way and with the use of one indicator. The method of construction of such indicators does not have to be the same for different sectors of economy, it is also defined in a different way by different countries and institutes. Leading indicators are used for the best forecasting of the future situation. Searching for such indicators is not easy in the countries with the stabilised economy as well as — in greater part — in countries where this situation changes quickly because of non-economic reasons. That is why two sessions of the conference were intended to the presentation of the analytical works on leading and composite indicators, and the following four — on the presentation of new methods and new indicators.

In this part there was only one paper from Poland, written by Joanna Klimowska, who presented composite economic sentiment indicator (A *Composite Economic Sentiment Indicator for Poland*).

Team of the researchers from Ifo — Chr. Hott, A. Kunkel and G. Nerb considered possibilities of identification of turning points of economic cycles with the use of business climate indicator used by this institute (*The Calculation of Turning Point Probabilities with the Ifo Business Climate*). Detection of turning points was also the subject of the paper of Alberto Baffigi and Antonio Bassanetti from the *Bank of Italy* (*Turning-Point Indicators form Business Surveys. Real-time Detection for the EURO Area and its Major Member Countries*). Second paper from the *Bank of Italy* also referred to the international comparisons, in this case to G7 countries. It was presented by Roberto Golinelli and Giuseppe Parigi (*Using Monthly Indicators to Bridge-Forecast Quarterly GDP for the G7 Countries*). Claudia Cicconi from ISAE presented an indicator used for synthetic description of the Italian economy (*A Smooth Indicator Based on Sectoral ISAE Business-Survey-Data*).

The team of the *Swiss Institute of Business Cycle Research KOF*: Richard Etter and Sybille Gubeli as well as Franz-Josef Klein, the retired employee of the *European Commission*, elaborated paper on indicators for GDP of the European Union (*Construction of Composite Indicators, Coincident and Leading the GDP*

of the European Union. Analysis of Detailed Business and Consumer Survey Results).

Special surveys

Special subjects presented in the separate group comprised different kinds of surveys on economic situation, mainly in the East and Central European countries. There were papers in which situation in particular areas in several countries was compared as well as surveys on the specific behaviour of enterprises and economy of different countries.

In the first group there were two papers presented by the team M. Carmen Guisan, Eva Aguayo and A. David Carballlas from the *Spanish University in Santiago de Compostela*. One of them was on cycles of the economic growth in the Eastern and Central Europe in comparison with chosen countries of the European Union (*Economic Growth and Cycles in Poland, Hungary, Czech Republic, Slovakia and Slovenia: A Comparison with Spain, Austria and other EU Countries, 1950—2000*), the second — on the influence of tourism on the economy of those countries (*Economic Effects of the Tourism in Central and Eastern Europe: 1990—2002*).

There was a great diversity of papers presented in those sessions. It is enough to mention that there were such elaborations as Polish paper of Aleksandra Gaweł from the *Poznań Academy of Economics* on forecasts of the labour market based on results of the business tendency surveys (*Forecasting Labour Market Situation on the Basis of Business Survey Data in Poland*), presentation of Eugeniusz Gorzelak from WSE on situation in Polish agriculture (*Business Conditions in Agriculture*), paper of the team from *Institute for Market Economics*: Jacek Fundowicz, Krzysztof Łapiński, Marcin Peterlik and Bohdan Wyżnikiewicz on regional economic cycles in Poland (*Regional Differentiation of Business Cycles in Poland*), paper of Marco Calabro and Paolo Carnazza from the *Institute of Promotion of Industry* on immigrants undertaking economic activity in Italy (*The Phenomenon of Immigrant Entrepreneurs in Italy: an Ad Hoc Survey*) or the elaboration of Linette Ellis and Jenny Terwin from South Africa on influence of AIDS on economy of this country (*The Impact of HIV/AIDS on Selected Sectors of the South African Economy*).

Seminar of the European Commission and OECD

Traditionally CIRET Conference was preceded by one day seminar on business tendency surveys organized by European Commission (Directoriate General on Economic and Financial Affairs, responsible for business tendency surveys in member states) and OECD. Seminar was divided into three sessions.

The subject of the first part was the discussion on works undertaken by OECD aiming at creation of the internet site containing wide information on business and consumer tendency surveys. The first version of the site was presented by the representatives of OECD. The intention of the authors was to create the source of the wide information on methodology of business and consumer tendency surveys recommended by international institutions and on surveys conducted in particular countries, their methodology, definitions and indicators, questionnaires and databases. Additionally it is proposed to be the place where information on organised international meetings on business tendency surveys is gathered as well as a forum of exchange of experiences between different countries. Because of the size of this works and the fact that similar works were planned by European Commission and CIRET the participants of the meeting proposed that the discussion is continued in the wider group with participation of the representatives of the countries outside OECD.

Second session was dedicated to the planned preparation of the metadata base concerning business and consumer tendency surveys. Representatives of OECD proposed to use the existing OECD handbook where recommended standards and particular definitions are included as an introduction to the further discussions on the contents of such base. Also in this work the wider group of experts should participate and proposal of OECD should be regarded as a starting point to the construction of the final base.

Third session was conducted by the representative of the European Commission and aimed at definition of priorities of the activities of Commission and member states for the nearest months in the field of the business and consumer surveys. The proposed areas of activities comprise improvement of the response rate by the decrease of the response burden, seasonal adjustment in the business tendency surveys and harmonization of the way of communication with respondents. Those works will be conducted within created task forces gathering representatives of the concerned countries.

Further discussion on the above subjects took place on the annual meeting organized by the Commission in October 2004 and will be continued at the joined seminar of the Commission and OECD planned for autumn 2005.

Prepared by
Izabella Zagoździńska — Central Statistical Office of Poland

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1205—1210

XXIII Conference on Multivariate Statistical Analysis (MSA 2004) & Symposium on Statistics to Commemorate the 110th Anniversary of Jerzy Neyman's Birthday

Łódź, Poland, 8—10 November 2004

The MSA 2004 Conference was organized by the Chair of Statistical Methods (University of Łódź) and the Polish Statistical Association. This conference has been organized every year, since 1981. This year, the Conference on Multivariate Statistical Analysis was held with a Symposium to commemorate the 110th anniversary of the birth Jerzy Spława-Neyman's. In the conference and symposium, there were 85 participants from different university centers from Poland and Ukraine. There were 46 papers presented at 11 sessions.

The conference was opened by the Chairman of the Organizing Committee: Prof. Czesław Domański. The opening speech was given by the Rector of the University of Łódź Prof. Wiesław Puś and the Dean of the Faculty of Economics and Sociology Prof. Paweł Starosta.

The scientific programme of MSA 2004 was very broad. The contributed papers were devoted to many statistical problems, such as multivariate distributions, nonparametric methods of estimation and verification of hypotheses, factor analysis, cluster analysis, Monte Carlo analysis, Bayesian inference, regression analysis and application of statistical methods in social and economics researches.

Symposium on Statistics to Commemorate the 110th Anniversary of Jerzy Neyman's Birthday

Two plenary sessions of symposium were dedicated to Prof. Jerzy Spława-Neyman, one of great founders of modern statistics. Jerzy Neyman made large contribution to probability theory, hypotheses testing, confidence intervals, survey sampling and other areas of mathematical statistics. His work would make an impact on fields ranging from astronomy and agriculture through biology and weather to social insurance. The scientific achievements of Prof. Jerzy Neyman and memories of his life and work were presented at the symposium.

The first plenary session of symposium (c*hair: Prof. Bronisław Ceranka* (*University of Poznań*)) started with a talk given by *Ryszard Zieliński* (*Polish Academy of Sciences, Warsaw*) entitled "***Jerzy Neyman and methodology of***

mathematical statistics". In the presentation, the author stressed the importance of the scientific activity of Jerzy Neyman to the creation of the methodological foundations of the probability theory and statistical inference. The author concentrated on the problems of model interpretation, hypotheses testing and confidence intervals.

Jan Kordos **(**Warsaw School of Economics, Warsaw) presented the lecture "**Professor J. Neyman's contribution to theory and practice of sampling methods — some Polish connections**". He analyzed the development of the concept of representative sampling, starting with Kiaer's (1895) and ending just before the 1934 Neyman's paper. The author presented Neyman's contribution to the theory of sampling and discussed his famous 1934 paper in details. It is stressed that despite many contributions of others to the development of sampling, the Neyman's 1934 paper played a crucial role in turning the pure mathematics into real sampling plans for actual randomly selected large scale surveys.

At the next plenary session (*chair: Prof. Mirosław Szreder* (*University of Gdańsk*)) a paper entitled "***Jerzy Spława-Neyman — my memories***" was presented. *Mirosław Krzyśko* (*University of Poznań*) shared his memories about meetings with Professor Neyman, starting from his academic years.

The presentation closing the symposium was entitled "*Jerzy Spława-Neyman — pillar of the mathematical statistics of the XXth century***". Author *Czesław Domański* (*University of Łódź*) reported scientific achievements of Neyman, particularly his large contribution to the theory of statistical tests and confidence intervals. He stressed that Spława-Neyman's works caused that statistics became a mathematical discipline.**

XXIII Conference on Multivariate Statistical Analysis

At the next sessions of the conference the papers presented the latest achievements in the field of the theory of multivariate statistical analysis and its applications in marketing, finance, risk management and health services. Titles of the papers with the names of the authors are presented below.

8 November 2004:

Plenary Session 3:

Chair: Prof. Krzysztof Jajuga (*University of Economics, Wrocław*)

- The Main Estimators of General Statistical Analysis, *Vyacheslav Girko* (*Ukrainian National Academy of Sciences, Kiev*)
- A procedure of division of means of three normal distributions into ε-homogeneous groups, *Stanisław Jaworski and Wojciech Zieliński* (*University of Agriculture, Warszawa*),

- Probability of a fuzzy event with t-norm, *Tadeusz Gerstenkorn and Joanna Gersternkorn* (*University of Łódź*),
- Multivalued Markov Processes, *Grażyna Trzpiot* (*University of Economics, Katowice*).

Plenary Session 4:

Chair: Prof. Jan Kordos (*Warsaw School of Economics, Warszawa*)

- The role of statistics in survey design in the field of social sciences, *Mirosław Szreder* (*University of Gdańsk)*,
- Verification of hypotheses on regression model's parameters for complex samples**,** *Czesław Domański* (*University of Łódź*),
- About the chemical balance weighing design with correlated errors, *Bronisław Ceranka and Małgorzata Graczyk* (*University of Agricultural, Poznań*),
- Chemical balance weighing design with correlated errors based on the ternary balanced block designs, *Bronisław Ceranka and Małgorzata Graczyk* (*University of Agricultural,Poznań*).

9 November 2004:

Two parallel sessions.

Session 1a:

Chair: Prof. Marek Walesiak (*University of Economics, Wrocław*)

- A new versions of pairwise classification method, *Mirosław Krzyśko and Waldemar Wołyński* (*University of Poznań*),
- Effectiveness examination of chosen methods of defining the number of clusters in a data set, *Jerzy Korzeniewski* (*University of Łódź*),
- On feature selection and multiple model approach in discriminant analysis, *Eugeniusz Gatnar* (*University of Economics, Katowice*),
- Discrimination of symbolic objects, *Andrzej Dudek* (*University of Economics,Wrocław*),
- Decomposition of classification error methods and their properties, *Dorota Rozmus* (*University of Economics, Katowice*).

Session 1b:

Chair: Prof. Janusz Wywiał (*University of Economics, Katowice*)

- Remarks on robust methods in the analysis of income distribution in Poland, *Alina Jędrzejczak* (*University of Łódź*),
- The average price dynamics — continuous time deterministic model, *Jacek Białek* (*University of Łódź*),

- Estimation of seasonal indices of deaths in Poland based on standardized data **,** *Zofia Mielecka — Kubień* (*University of Economics, Katowice*),
- Some considerations on population premature mortality (on the basis of analysis of mortality in province Lodzkie and in the city of Lodz in the years 1990—2000), *Joanna Majczak* (*University of Economics, Wrocław*),
- The statistical analysis of socio-economic characteristics on the example of communes of Podkarpackie province, *Andrzej Mantaj, Wiesław Wagner* (*University of Information Technology and Management, Rzeszów*).

Session 2a:

***Chair: Prof. Ryszard Zieliński* (*Polish Academy of Sciences, Warszawa*)**

- Estimation of bias and variance of sample median by jackknife and bootstrap methods, *Krystyna Pruska* (*University of Łódź*),
- Goodness-of-fit tests for ROC curves, *Agnieszka Rossa* (*University of Łódź*),
- Join monitoring of the mean, dispersion and asymmetry in process control, *Grzegorz Kończak* (*University of Economics, Katowice*),
- On the phase transitions in the Kendall's shape space, *Daniel Kosiorowski* (*University of Economics, Kraków*),
- Multiple endpoints, *Dariusz Parys* (*University of Łódź*).

Session 2b:

***Chair: Prof. Wojciech Zieliński* (*University of Agriculture, Warszawa*)**

- Analysis of the linear regression model by divisibled submatrix of the system, *Wiesław Wagner* (*Rzeszów*) *and Anna Budka* (*University of Agriculture, Poznań*),
- Some nonparametric estimators of regression function, *Aleksandra Baszczyńska* (*University of Łódź*),
- MART – multiple additive regression trees and their applications, *Joanna Trzęsiok* (*University of Economics, Katowice*).
- Prediction of time series using jackknife method, *Janusz Wywiał and Tomasz Żądło* (*University of Economics, Katowice*).
- GARCH models of time series on DAM, *Alicja Ganczarek* (*University of Economics, Katowice*).

Session 3:

Chair: Prof. Mirosław Krzyśko (*University of Poznań*)

- On linear sufficiency in the general linear model with additional information, *Paweł Pordzik* (*University of Agriculture, Poznań*),
- An influence of correlation among variables on efficiency of the modified synthetic estimator: a Monte Carlo analysis, *Tomasz Jurkiewicz* (*University of Gdańsk*),

- Comparison of various prior estimation methods applied to hierarchical Bayes estimation, *Jan Kubacki* (*Statistical Office*, *Łódź*),
- Bootstrap confidence regions based on the Mahalanobis depth measure of two-dimensional samples, *Małgorzata Kobylińska* (*University of Warmia and Mazury, Olsztyn*) *and Wiesław Wagner* (*University of Information Technology and Management, Rzeszów*),
- Kriging — a method of static interpolation spatial data, *Jan Kowalik* (*Technical University of Częstochowa*).

Session 4:

Chair: Prof. Wiesław Wagner (*University of Information Technology and Management, Rzeszów*)

- Latent class models applications in the research of consumer preferences, *Andrzej Bąk* (*University of Economics, Wrocław*),
- On some modification of support vector machines, *Michał Trzęsiok* (*University of Economics, Katowice*),
- Dynamic modelling of values of accounting units of OPF using ARIMA models, *Agnieszka Orwat and Grażyna Trzpiot* (*University of Economics, Katowice*).

10 November 2004: Session 1:

Chair: Prof. Krystyna Pruska (*University of Łódź*)

- Effective portfolio construction as multivariate classification problem, *Urszula Skórnik — Pokarowska* (*University of Agriculture, Warszawa*), *Arkadiusz Orłowski* (*Polish Academy of Science, Warszawa*),
- Chosen methods of effective constructions and admissible stock portfolios, *Adam Depta* (*University of Łódź*),
- On the certain collective risk model for non-homogenous populations, *Jolanta Grala* (*University of Poznań*),
- Making decision in risk, *Agata Szczukocka* (*University of Łódź*),
- Implementation of simplex algorithm for risk limitation using the example of agronomic enterprise activity, *Marek Rosolak and Konrad Zarzecki* (*University of Łódź*),
- The rough set approach for multicriteria classification problem– the statistical characteristics of decision rules, *Iwona Gruszka* (*University of Economics, Katowice*),
- Assessment of forecasts treated as random variables, *Ewelina Zaleska* (*University of Łódź*).

The papers presented at the conference were interesting and involved many important statistical problems, both from the theory and application point of view.

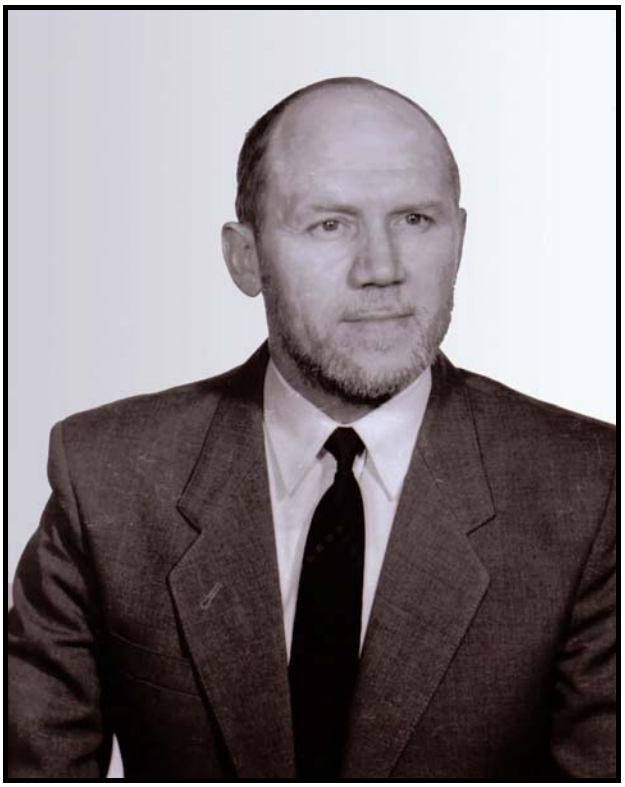
The papers will be published in "Acta Universitatis Lodziensis — Folia Oeconomica".

The next Multivariate Statistical Analysis conference will be held in November 7-9, 2005 in Łódź. Scientists interested in this Conference are kindly requested to send their application to the Scientific Secretary of MSA 2005 at the following address: Anna Witaszczyk, 24th Conference on Multivariate Statistical Methods, Chair of Statistical Methods, University of Łódź, 90-214 Łódź, Rewolucji 1905 r. nr 41, phone 0-42 635-53-35, e-mail : msa@uni.lodz.pl .

Prepared by Dorota Pekasiewicz

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1211—1213

OBITUARY



PROFESSOR A. RYSZARD WÓJCIK
(1937-2003)

On December 11, 2003, Dr. Aleksander Ryszard Wójcik died at his home in Toluca, Mexico. He was 66 years old. Following his wishes his remains have been brought back to Poland and he is now resting in Nadarzyn near Warsaw. He was born on August 4, 1937 in Brzezienko Nowe, Poland.

His life was deeply affected by the World War II during which he lost his father, and which provided a rough start for his formal education. Still, he obtained his first Masters degree in economics from the Faculty of Statistics at the Central School of Planning and Statistics (SGPiS — now the Warsaw School of Economics) in 1958 at the age of 21. He then pursued a second degree, in mathematics, at the Faculty of Mathematics and Physics of the Warsaw

University, lecturing and working simultaneously towards his doctorate at the Warsaw Agricultural University (SGGW). He obtained his Master degree in mathematics in 1964 and PhD degree in 1965. He then began his work as a full time professor and researcher at the SGGW, where he obtained post-doctoral habilitation with specialisation in experimental design in 1981.

During the 60's, 70's and 80's Dr. Wójcik applied his extensive knowledge of statistical methods in various data analyzes. In addition to functioning as a faculty member at the SGGW, he worked at the Drug Institute in Warsaw, the Institute for Potato Research in Bonin and the Agricultural University in Siedlce, and engaged in multiple collaborations in areas of medicine, forestry, genetics and agriculture. During this period he was an author and co-author of over 30 scientific papers and 5 books. In 1984 Dr. Wójcik further expanded his research while spending 5 months as visiting professor at the University of California in Berkeley.

Dr. A.R. Wójcik concentrated mainly on the application of mathematical and statistical methods to biological and agricultural sciences. At the beginning of his scientific career he worked on statistical modeling of drug testing. Since 1965 his research work concerned mostly linear modeling of agricultural experiments (particularly long term experiments). Dr. Wójcik also worked on the application of multiple regression in assessing the influence of dynamical processes of vegetation period on characteristics of agricultural plants.

Dr. A.R. Wójcik cooperated with the Central Statistical Office of Poland (CSO). From 1966 to 1982 he was a member of the Mathematical Commission of the CSO, being actively involved in designing agricultural sample surveys. He was a member of the Polish Biometrical Society, the Polish Mathematical Society, the Polish Statistical Association, and the Bernoulli Society. It is worth emphasizing that he contributed to the Encyclopedia of Statistics as an author.

He was also well recognized as a teacher. He delivered many lectures to students of various faculties of the SGGW. His textbooks „*Mathematical statistics with elements of probability*", „*Methods of agricultural experiments*" (with J. Trętowski), and „*Experimental design and statistical inference in biometry*" (with Z. Laudański) are still popular among students and young researchers.

In 1987 Dr. Wójcik moved to Mexico, starting as a visiting professor at the Agricultural University of Chapingo. In 1988 he undertook a full time professor position at the Monterrey Technological Institute (ITESM) in Toluca where he continued to work until heath problems forced him to retire in 2003. During that time he published more than 20 scientific articles and 3 textbooks in the areas of Mathematics, Statistics and Experimental Design. His latest research was focused mostly on evaluating and modelling the trends of air pollution in Mexico, a topic of great importance given the health and environmental risks that are increasingly being associated with the high levels of ozone and CO emissions produced by North America. Using bootstrap methods and transfer function methodology his research group designed models for detection of trends in air pollution levels and

ozone forecasting. The relevance of this work was confirmed by the vast financial support granted by the Mexican Science and Technology Council (CONACYT) and the numerous international publications and meetings.

In addition to being a well known scientist, Dr. Wójcik took great pride and joy in teaching. During his 15 years of work as a professor at the ITESM he taught over 100 courses in the areas of Statistics, Mathematics, Probability, Linear Modelling, Regression Analysis, Experimental Design and Statistical Quality Control to thousands of undergraduate and graduate students. He also supervised several Master and PhD theses. During summer terms Dr. Wójcik taught Statistics as visiting professor at Southern Illinois University (Edwardsville, USA) in 1993 and in Quito, Ecuador in 1998. His dedication, profound expertise and true concern about transmitting knowledge to the students earned him the highest appreciation in form of the „*Golden Lamb"* best professor prize in 1992 and "*Recognition to Excellence*" price in 2003. He was always committed to transmitting not only statistical knowledge but also every-day values such as honesty, integrity and responsibility. He is deeply cherished by generations of graduates.

A.R. Wójcik is survived by his wife Jadwiga and two children, Ewa Joanna and Aleksander Karol, who are both pursuing PhD degrees in the USA. Despite his strongest desire and efforts he did not meet his first grandchild — Emilia Aleksandra was born exactly 6 months after the death of her grandfather.

Even while facing continuous deterioration of his health, Dr. Wójcik kept devoting his days to Statistics. In 2003, together with his son, he finished writing his last book „*Applied Probability and Statistics*", which is soon to be published in electronic version. In recognition of his dedication and contributions as scientist, teacher and friend, a teaching hall bearing his name is to be inaugurated at the ITESM Campus Toluca this winter.

Professor A. R. Wójcik was an Associate Editor of our journal *Statistics in Transition* since 1993 and significantly contributed to its development.

He was a passionate educator, prominent scientist and caring husband, father and friend. He is missed dearly. Always.

STATISTICS IN TRANSITION, December 2004
Vol. 6, No. 7, pp. 1215—1216

Acknowledgements

Referees of Volume 6

The Editorial Board wishes to thank the following referees who have generously given their time and skills to the Statistics in Transition during the period from January 2003 to December 2004, i.e. for preparing Volume 6.

Czesław Bracha, Warsaw School of Economics, Warsaw, Poland.

Stig Danielsson, Linköping University, Sweden.

Czesław Domański, University of Łódź, Łódź, Poland

Sławomir Dorosiewicz, Warsaw School of Economics, Warsaw, Poland

Ewa Frątczak, Central Statistical Office of Poland, Warsaw, Poland.

Wayne A. Fuller, Iowa State University, USA

Elżbieta Gołata, University of Economics, Poznań, Poland.

Marek Góra, Warsaw School of Economics, Warsaw, Poland

Marek Gruszczyński, Warsaw School of Economics, Warsaw, Poland

Johan Heldal, Statistics Norway, Norway

Anders Holmberg, Statistics Sweden, Sweden

Krzysztof Jajuga, Wrocław University of Economics, Wrocław, Poland

Sven-Erik Johansson, Karolinska Medical University, Sweden

Graham Kalton, WESTAT, Inc. Washington, USA

Irena Kasperowicz-Ruka, Warsaw School of Economics, Warsaw, Poland.

Jan Kordos, Warsaw School of Economics, Warsaw, Poland.

Irena Kotowska, Warsaw School of Economics, Warsaw, Poland

Liliana Kursa, Central Statistical Office, Warsaw, Poland

Seppo Laaksonen, Statistics Finland

Janis Lapins, Bank of Latvia, Riga, Latvia

Risto Lehtonen, Jyväskylä University, Finland

Małgorzata Misztal, , University of Łódź, Łódź, Poland

Marek Męczarski, Warsaw School of Economics, Warsaw, Poland.

Wojciech Niemiro, Warsaw University, Warsaw, Poland

Lucyna Nowak, Central Statistical Office of Poland, Warsaw, Poland.
Jerzy Nowakowski, Warsaw School of Economics, Warsaw, Poland.
Paul Ollila, Statistics Finland, Finland
Tomasz Panek, Warsaw School of Economics, Warsaw, Poland.
Jan Paradysz, University of Economics, Poznań, Poland
Dariusz Parys,, University of Łódź, Łódź, Poland
Dorota Pekasiewicz, University of Łódź, Łódź, Poland
Richard Platek, Formerly Statistics Canada, Ottawa, Canada
Jarosław Podgórski, Warsaw School of Economics, Warsaw, Poland.
Waldemar Popiński, Central Statistical Office, Warsaw, Poland
Krystyna Pruska, University of Łódź, Łódź, Poland
J.N. K. Rao, Dept.of Maths and Statistics, Carleton University, Canada
Teresa Słaby, Warsaw School of Economics, Warsaw, Poland.
Honorata Sosnowska, Warsaw School of Economics, Warsaw, Poland
Adam Szulc, Warsaw School of Economics, Warsaw, Poland
Mirosław Szreder, University of Gdańsk, Gdańsk, Poland
Daniel Thorburn, Stockholm University, Sweden
Imbi Traat, Tartu University, Estonia
Vijay Verma, Consultant in Survey Methodology, India
Jacek Wesołowski, Warsaw University of Technology, Warsaw, Poland
Robert Wieczorkowski, Central Statistical Office, Warsaw, Poland
Janusz Witkowski, Central Statistical Office, Warsaw, Poland
Janusz Wywiał, Academy of Economics, Katowice, Poland
Jan Wretman, Stockholm University, Sweden
Aleksander Zeliaś, Cracov University of Economics, Cracov, Poland
Li Chun Zhang, Statistics Norway, Norway